Neuro Inspired Computing in FLASH

2014 Neuro Inspired Computational Elements Workshop

Kevin Gomez
SSD Architecture

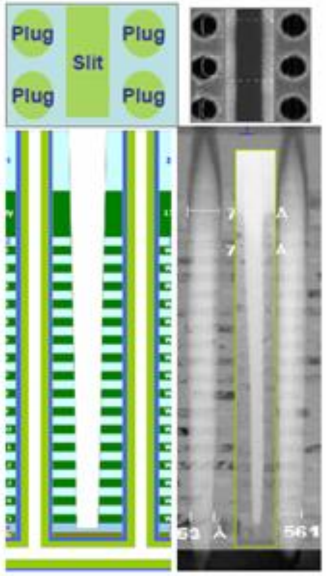
Summary

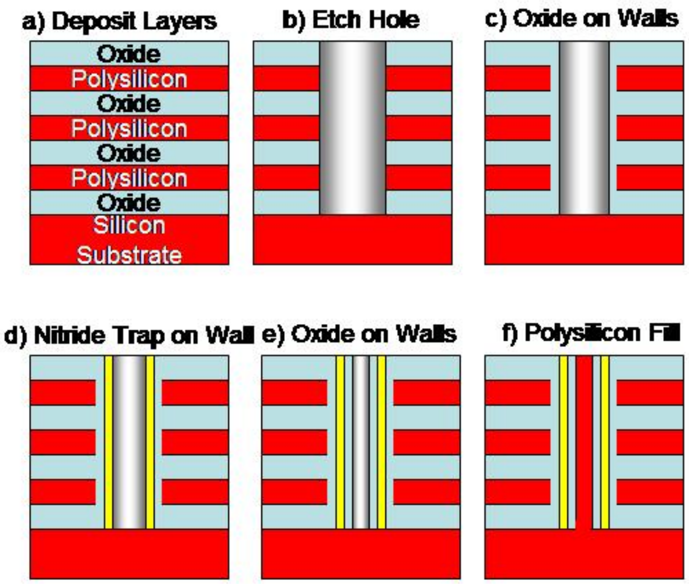
- Embed specialized hardware into Solid State Drives (SSDs) for cortical processing assist
- Architecture modeling shows ~100x lower J/op compared to processing on host
- Identical to standard SSD - same manufacturing process and cost
- Re-purposed as cortical processor through firmware
- Open standards, e.g. same APIs as GPGPU, OpenCL, PyNN

Seagate

NAND Flash

has successfully transitioned to 3D





Jim Handy, 2013

An ingenious breakthrough which enables multiple layers of memory **without needing to pattern each layer**
Decoupling NAND Flash production CAPEX from lithography (45% for planar down to 15% for 3D)*

*Samsung Analyst Day 2013 Memory Business

3D NAND Process

devices shipping in volume

3D NAND Process

devices shipping in volume

Silicon Substrate

3D NAND Process

devices shipping in volume

3D NAND Process
devices shipping in volume

3D NAND Process

devices shipping in volume

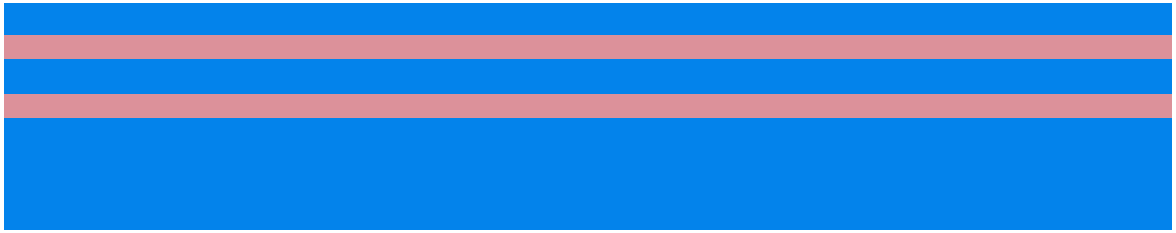
3D NAND Process

devices shipping in volume

Silicon Substrate

3D NAND Process
devices shipping in volume

Silicon Substrate



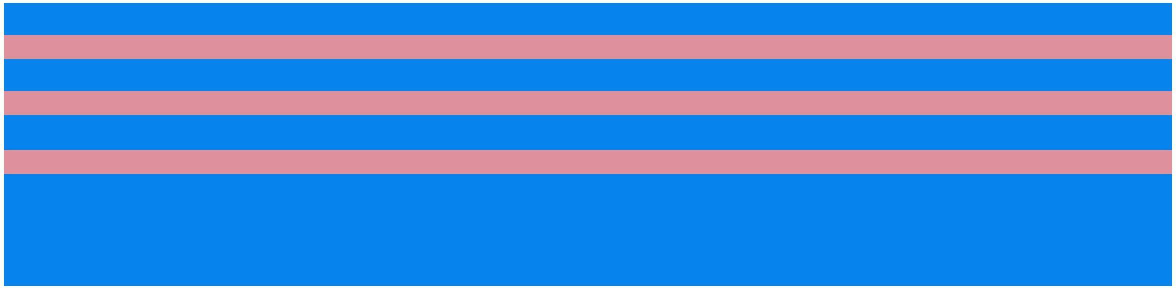
3D NAND Process

devices shipping in volume

Silicon Substrate

3D NAND Process

devices shipping in volume



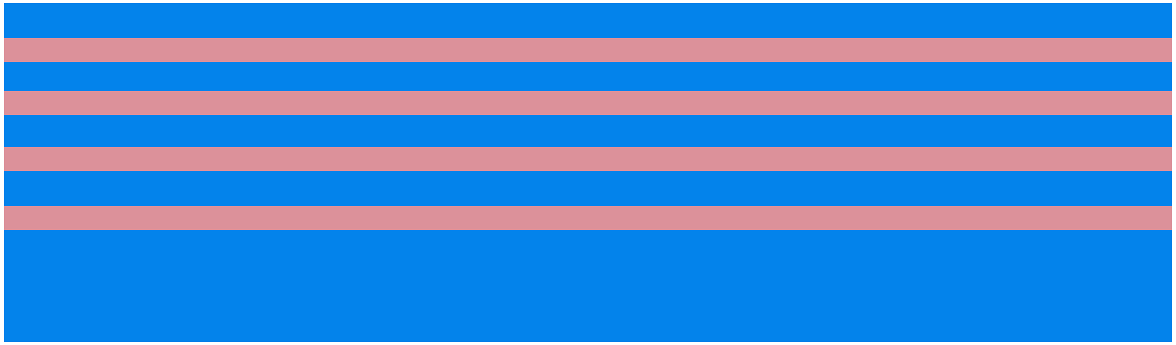
3D NAND Process

devices shipping in volume

Seagate

3D NAND Process

devices shipping in volume



3D NAND Process

devices shipping in volume

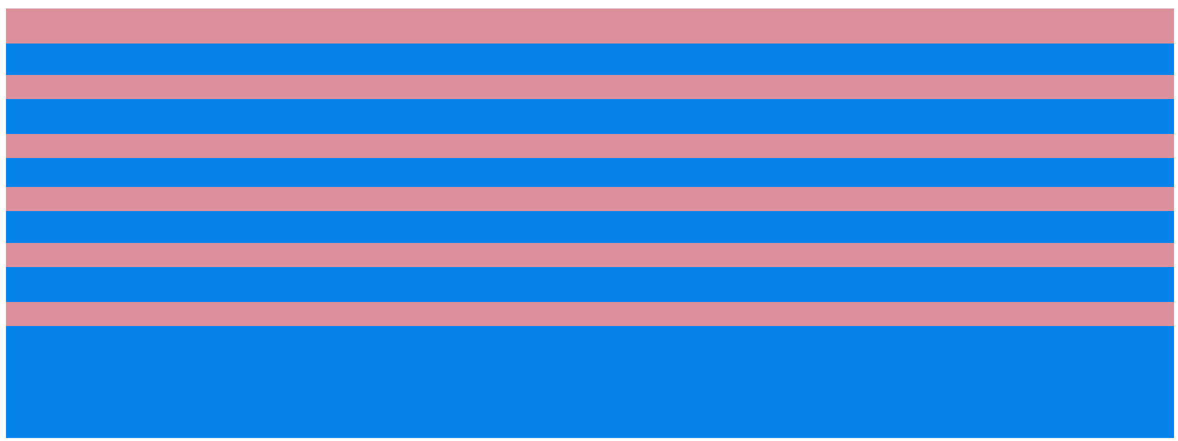
3D NAND Process

devices shipping in volume

Seagate

3D NAND Process

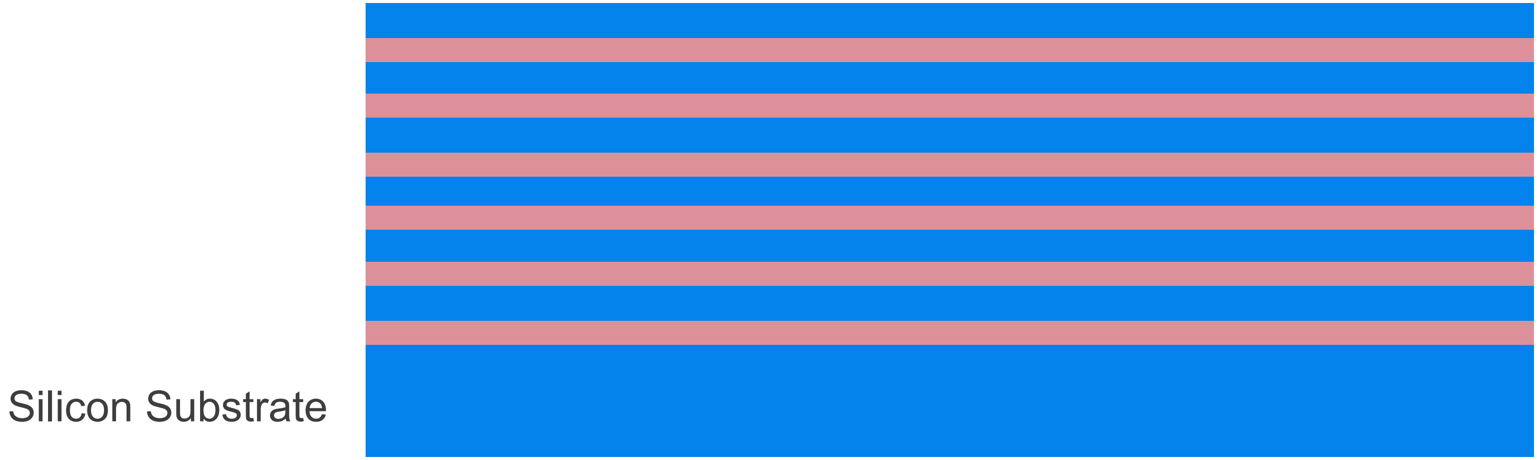
devices shipping in volume



Silicon Substrate

3D NAND Process

devices shipping in volume

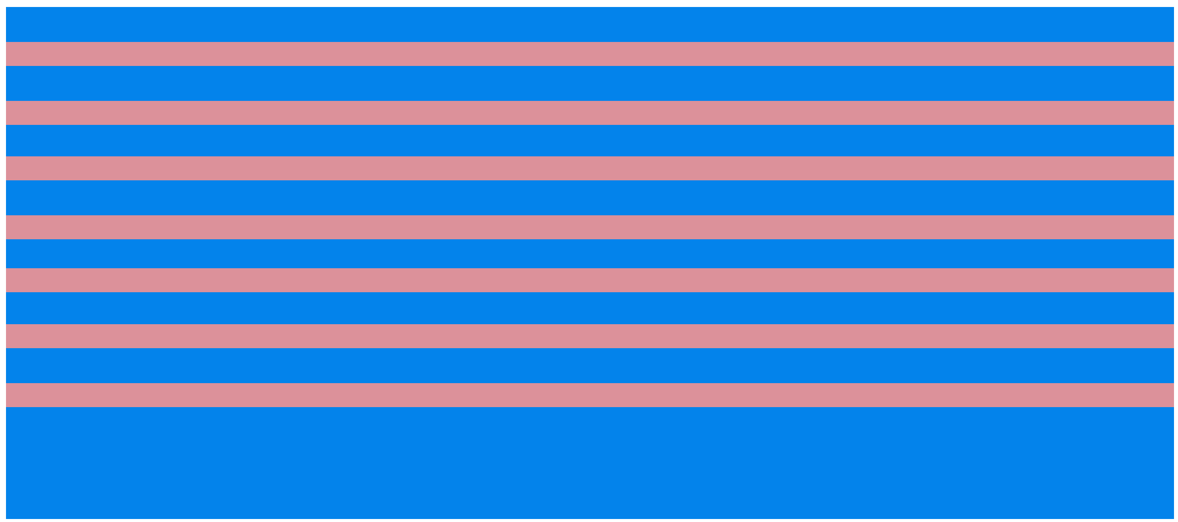


3D NAND Process

devices shipping in volume

3D NAND Process

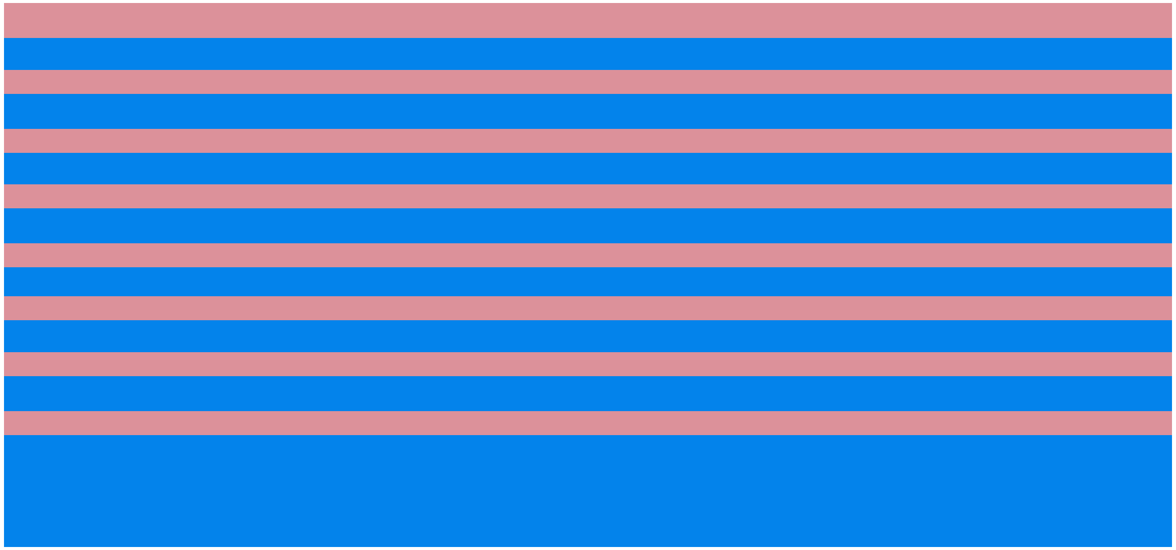
devices shipping in volume



Silicon Substrate

3D NAND Process

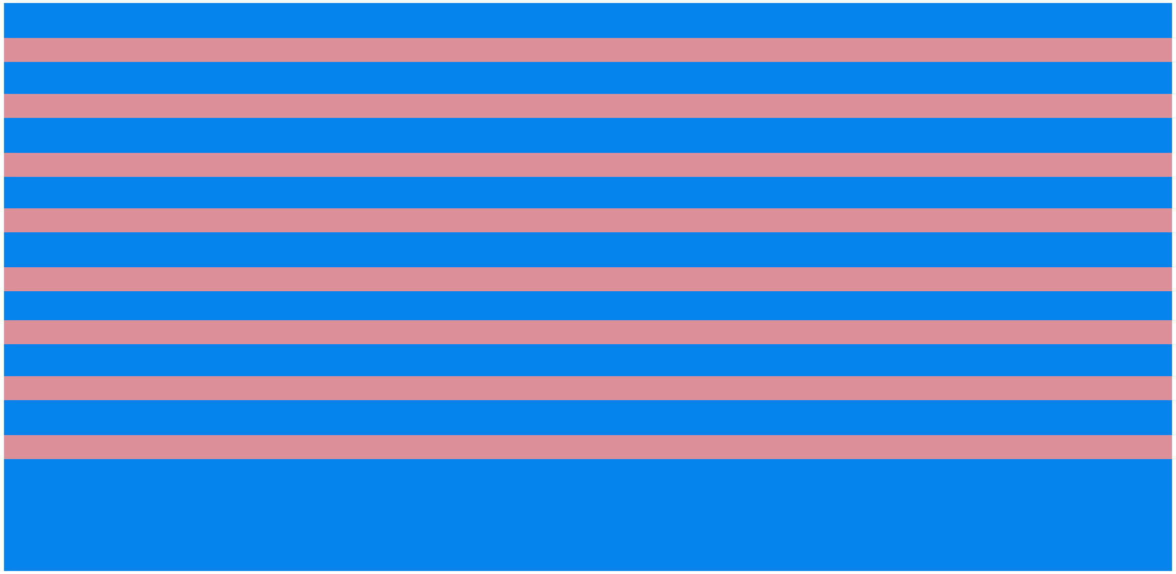
devices shipping in volume



Silicon Substrate

3D NAND Process

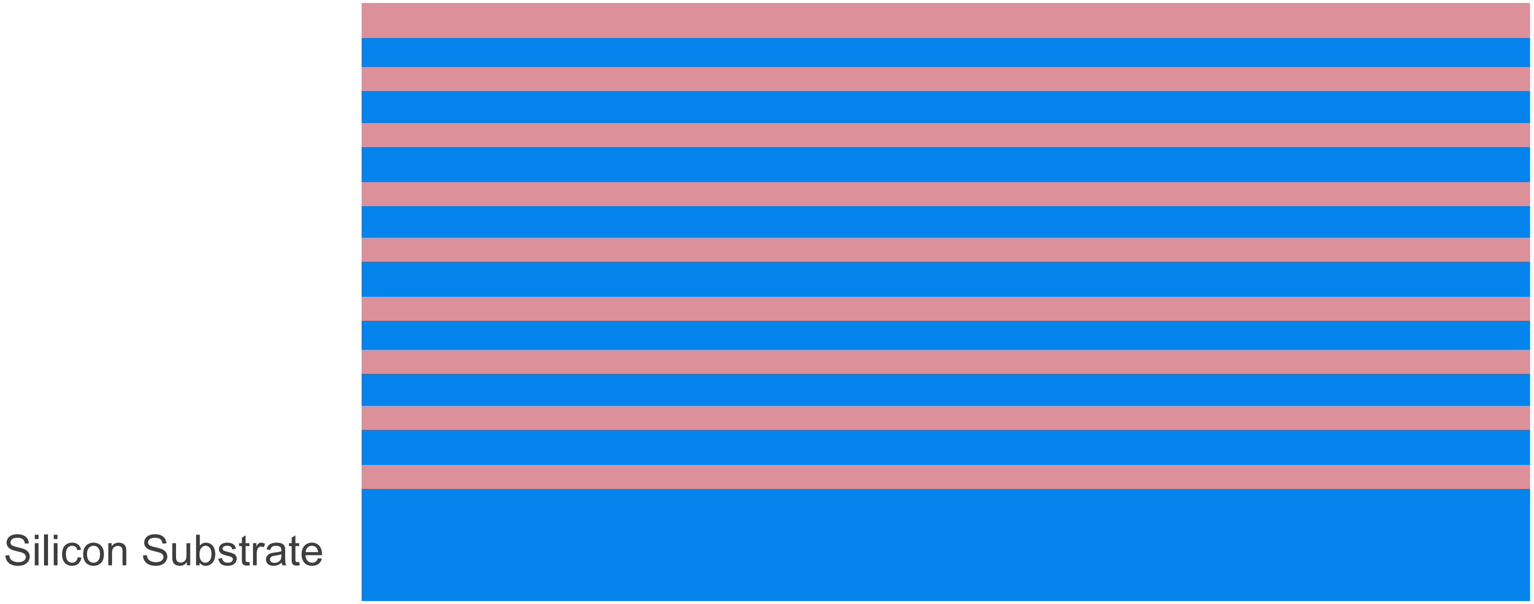
devices shipping in volume



Silicon Substrate

3D NAND Process

devices shipping in volume



Seagate

3D NAND Process

devices shipping in volume

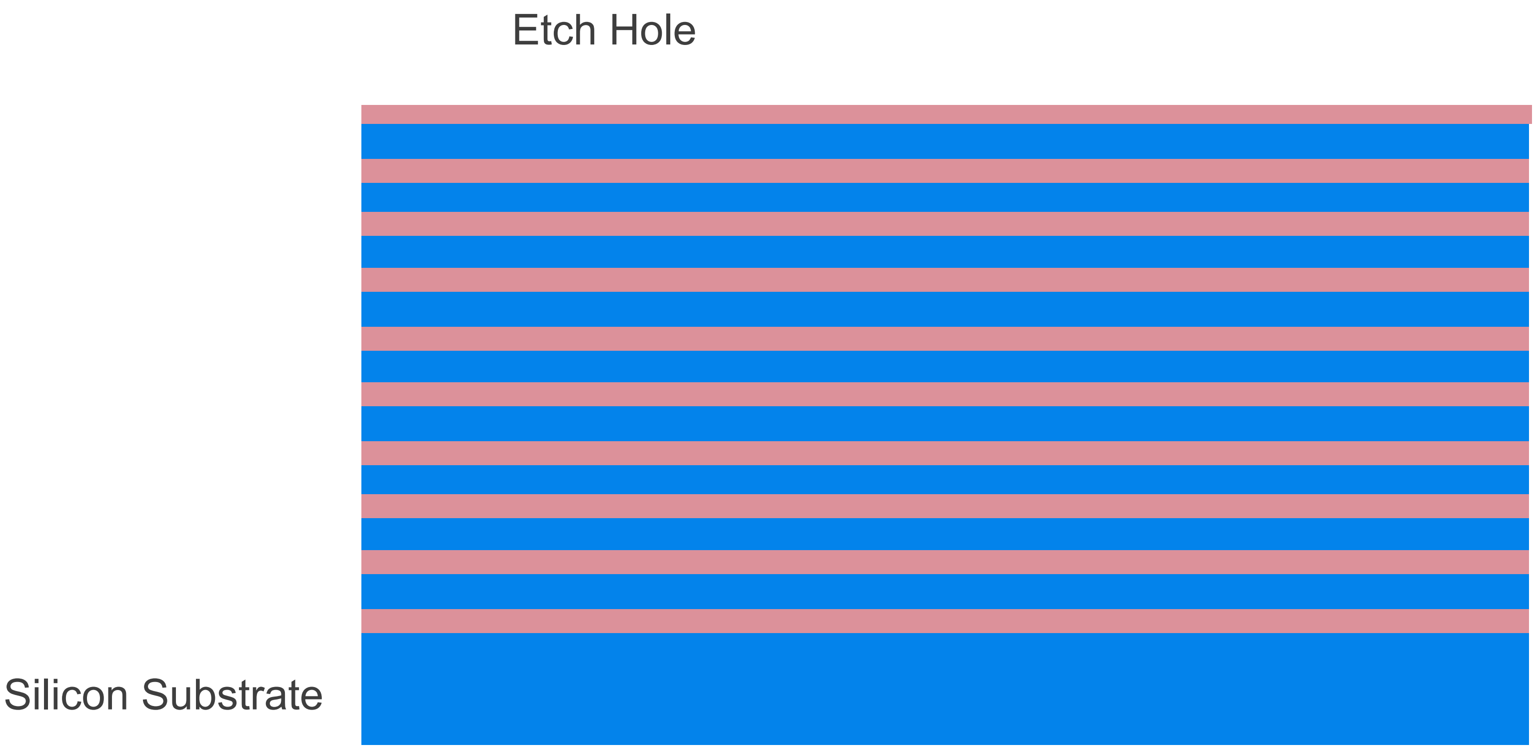
Silicon Substrate

3D NAND Process
devices shipping in volume

Silicon Substrate

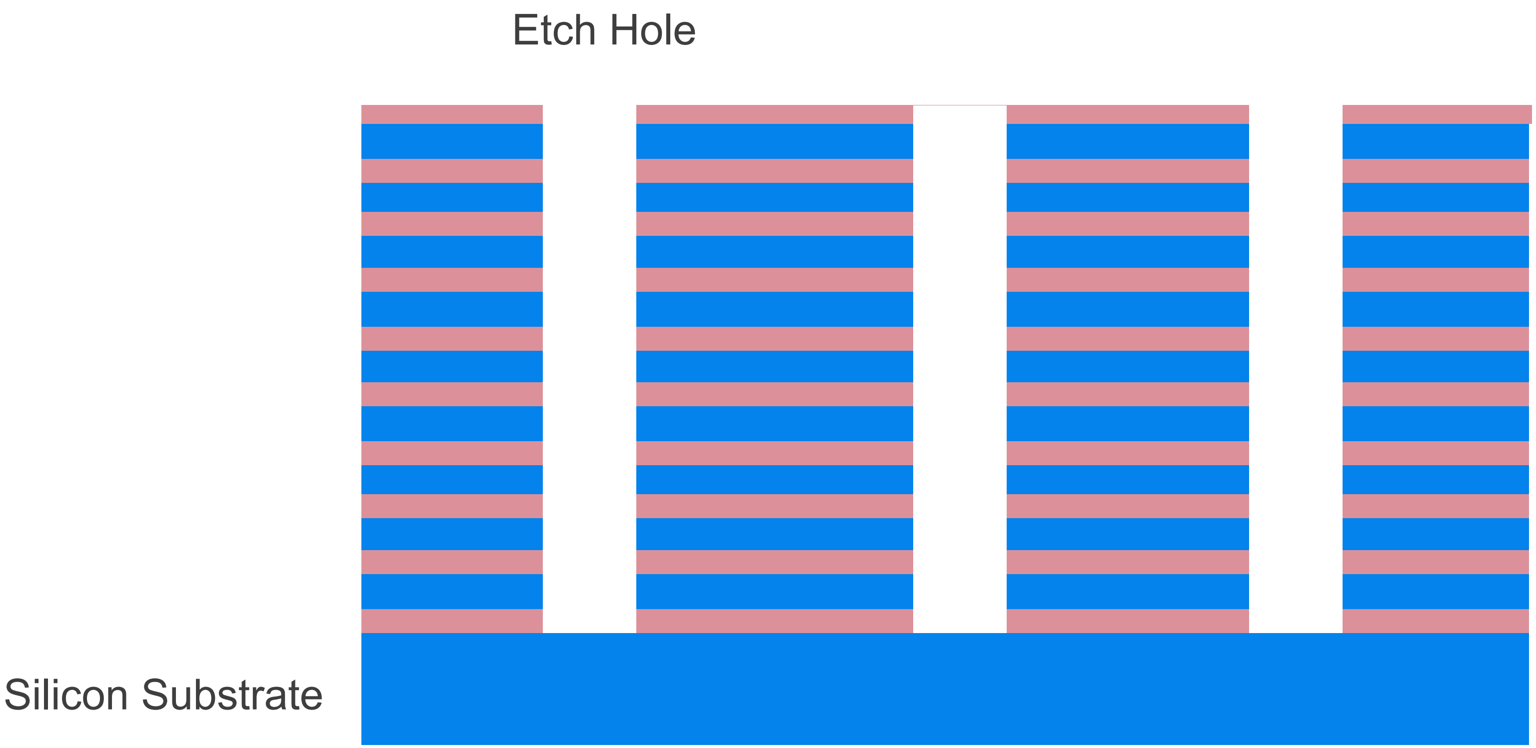
3D NAND Process

devices shipping in volume



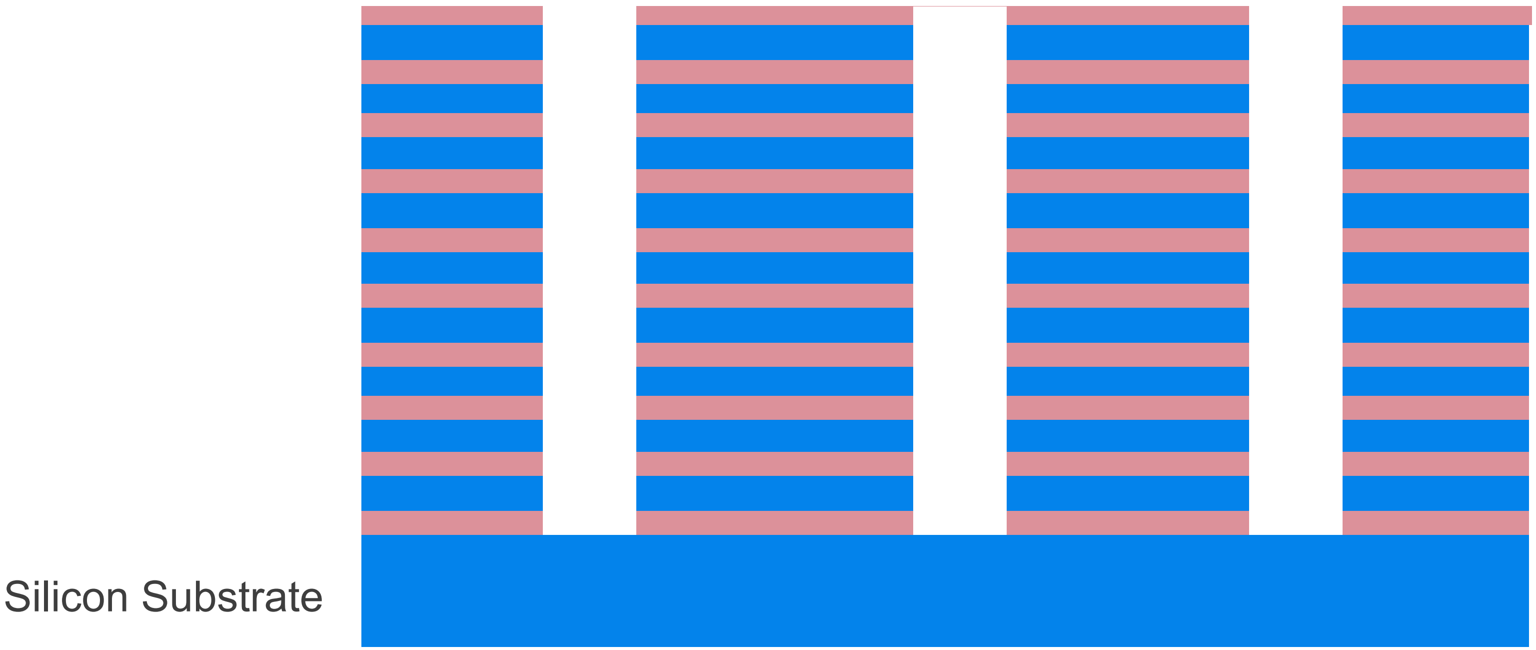
3D NAND Process

devices shipping in volume



3D NAND Process

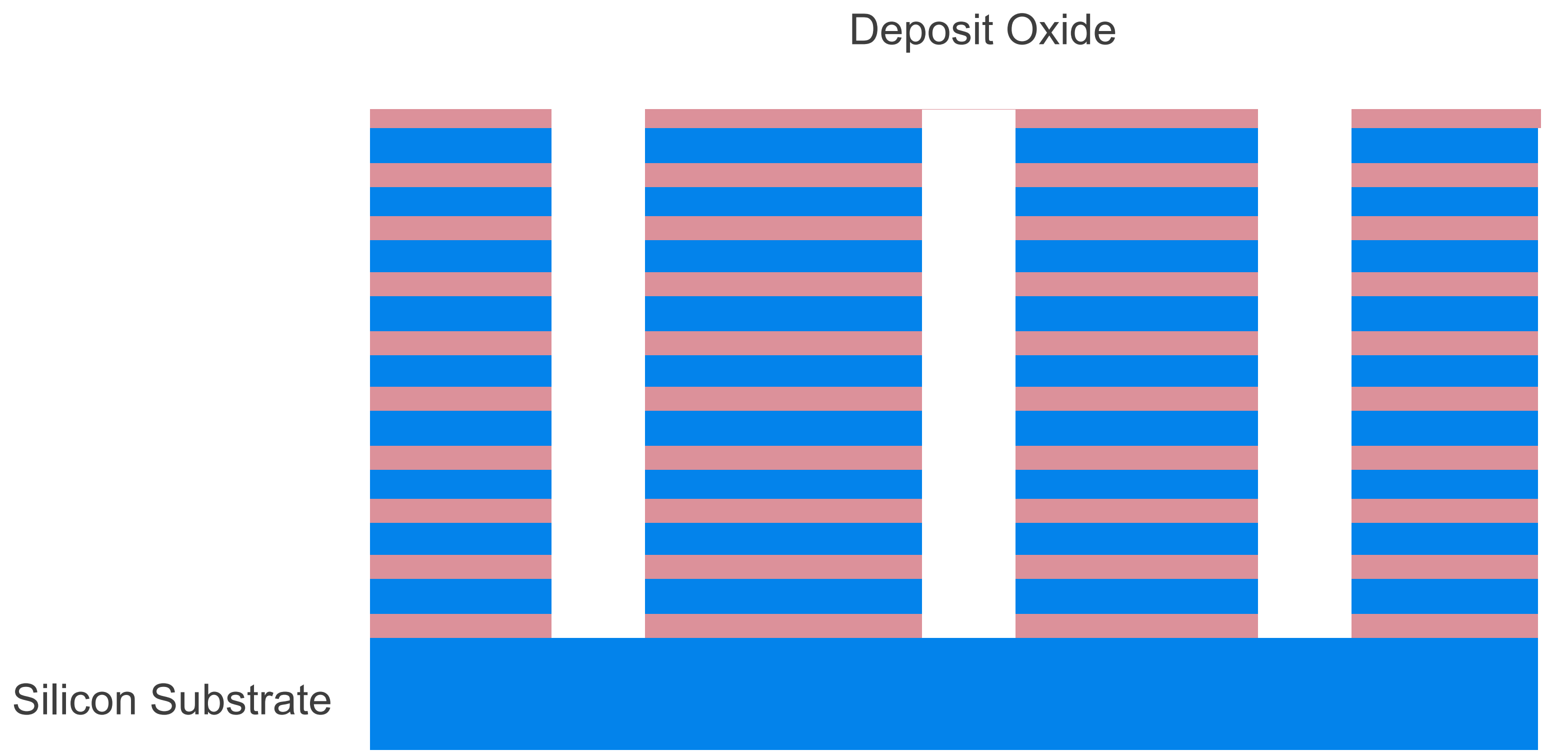
devices shipping in volume



Seagate®

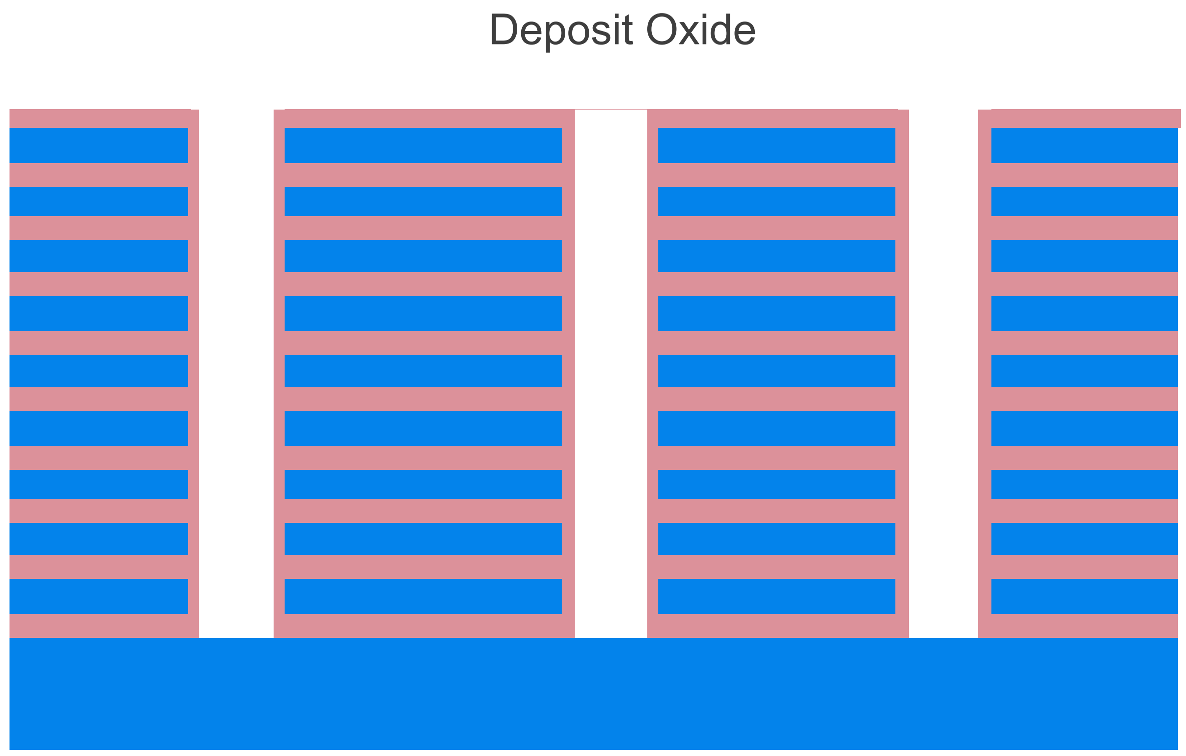
3D NAND Process

devices shipping in volume



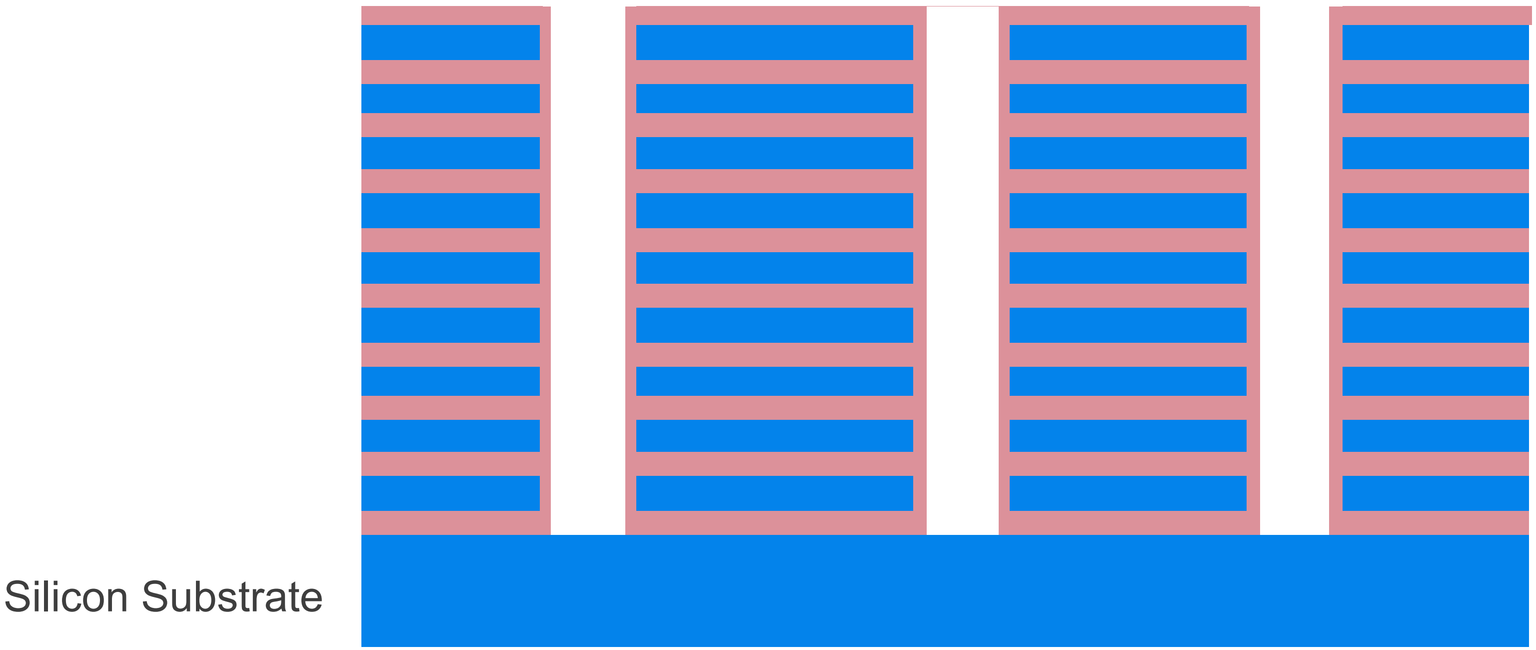
3D NAND Process

devices shipping in volume



Seagate

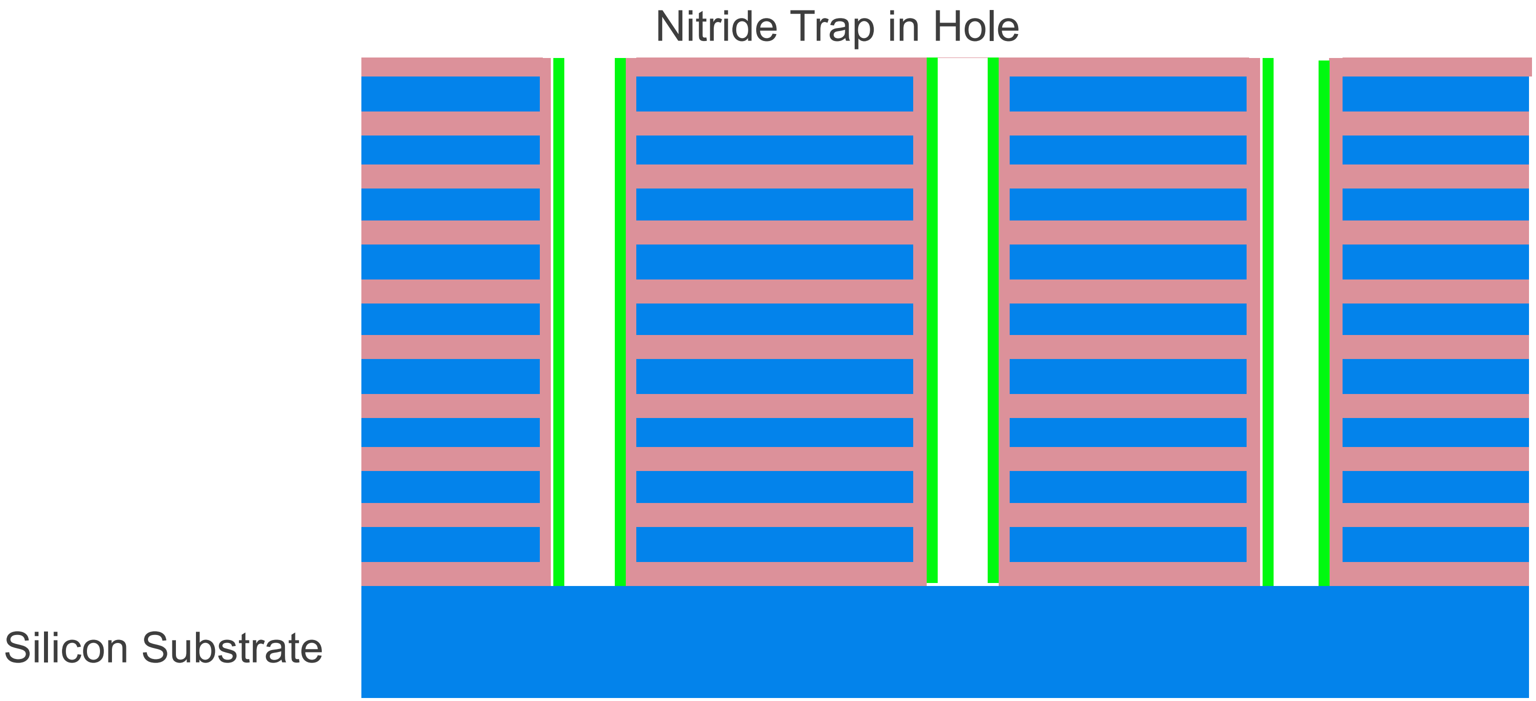
3D NAND Process
devices shipping in volume



Seagate®

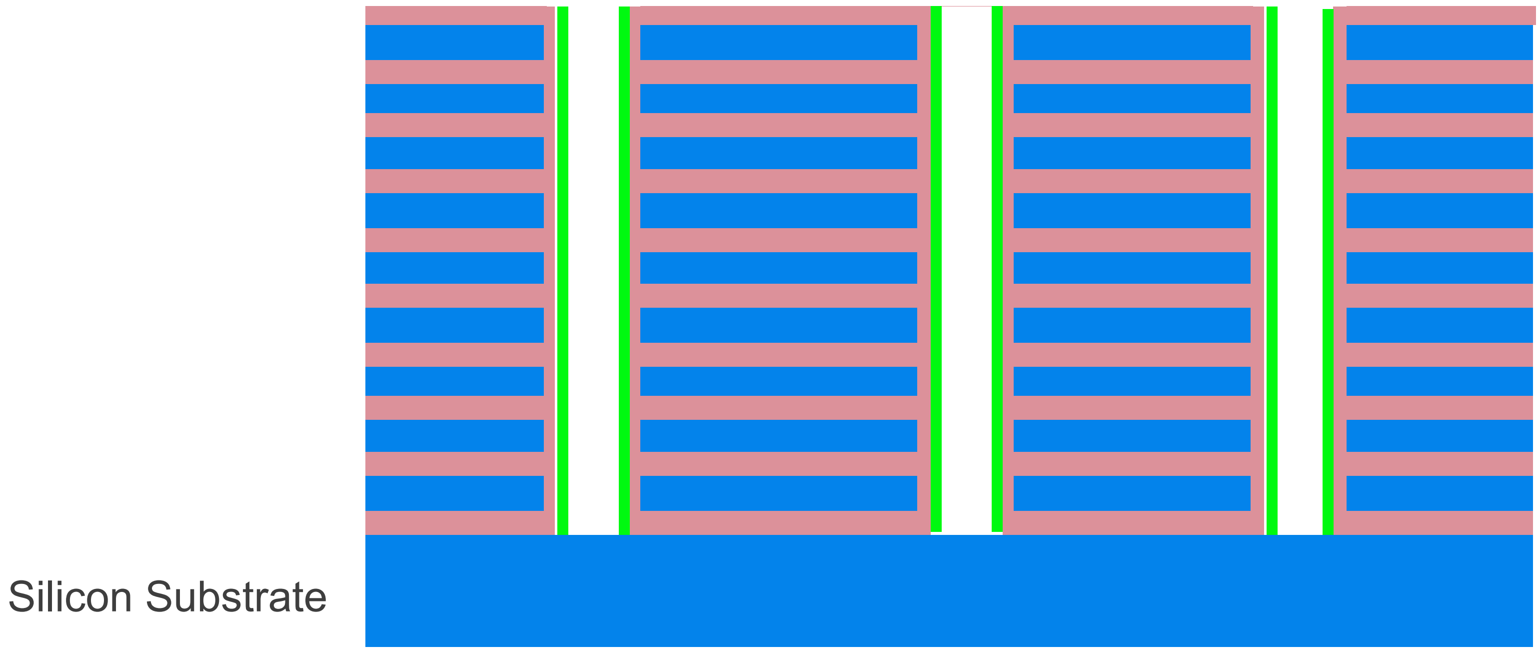
3D NAND Process

devices shipping in volume

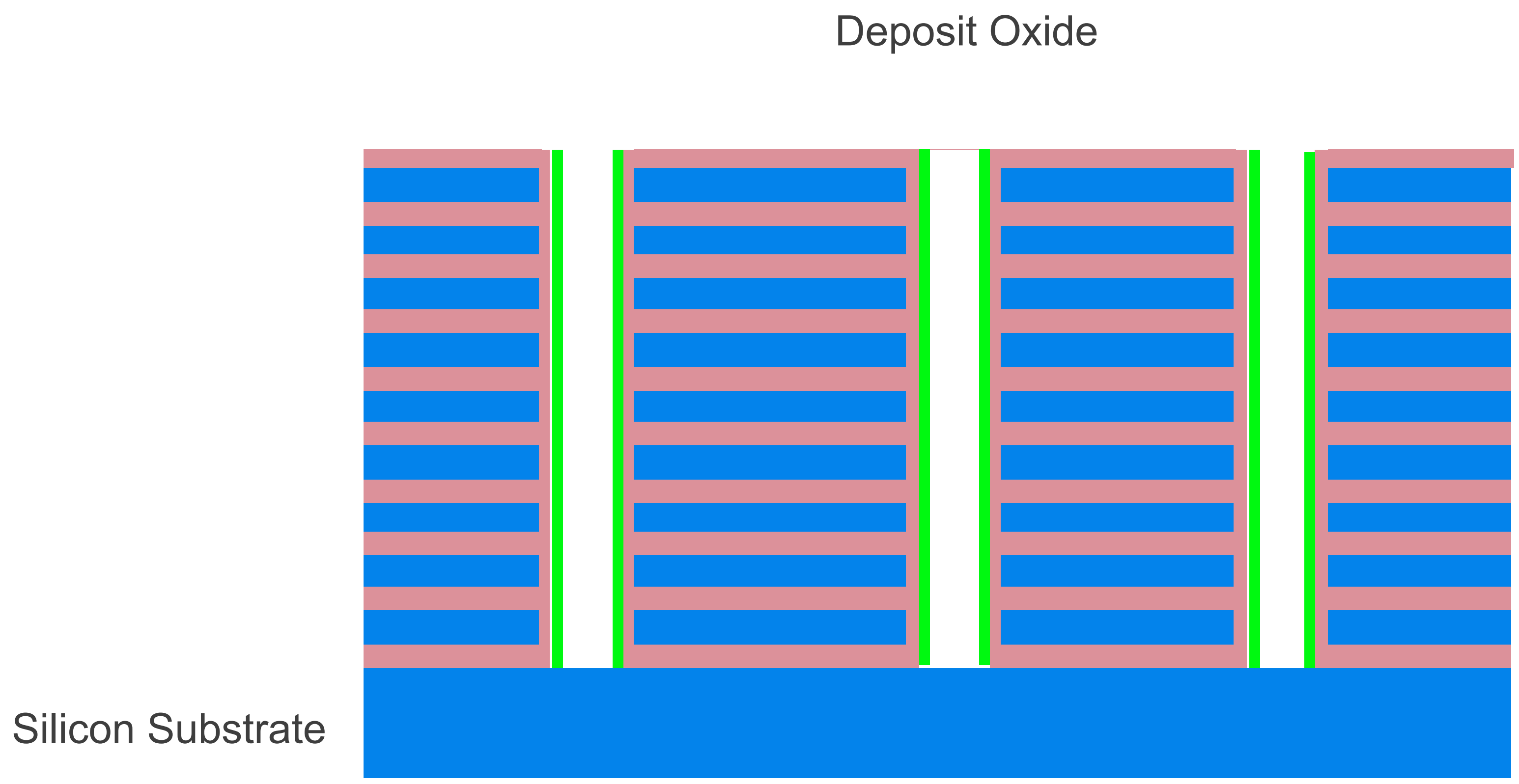


3D NAND Process

devices shipping in volume

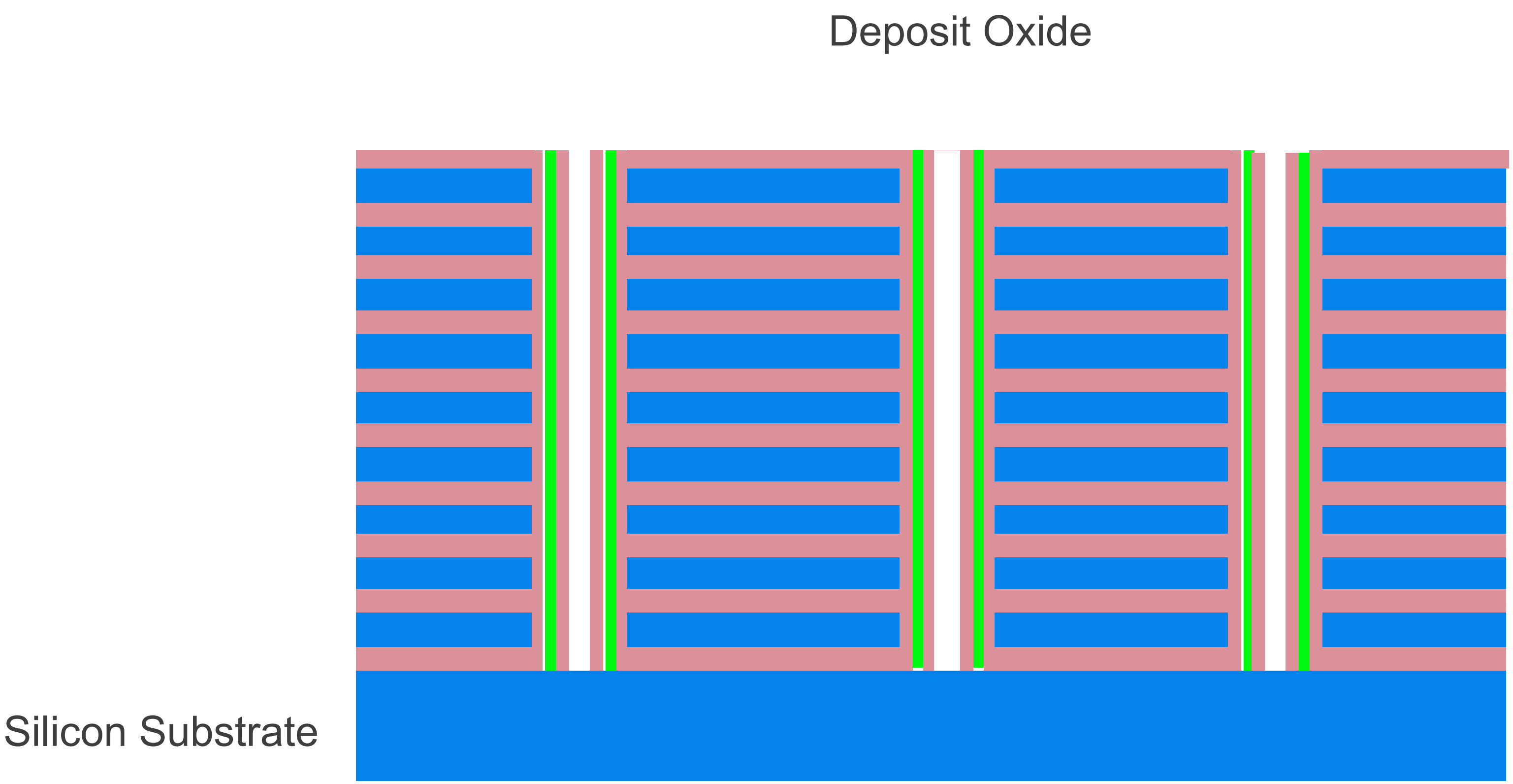


3D NAND Process
devices shipping in volume



3D NAND Process

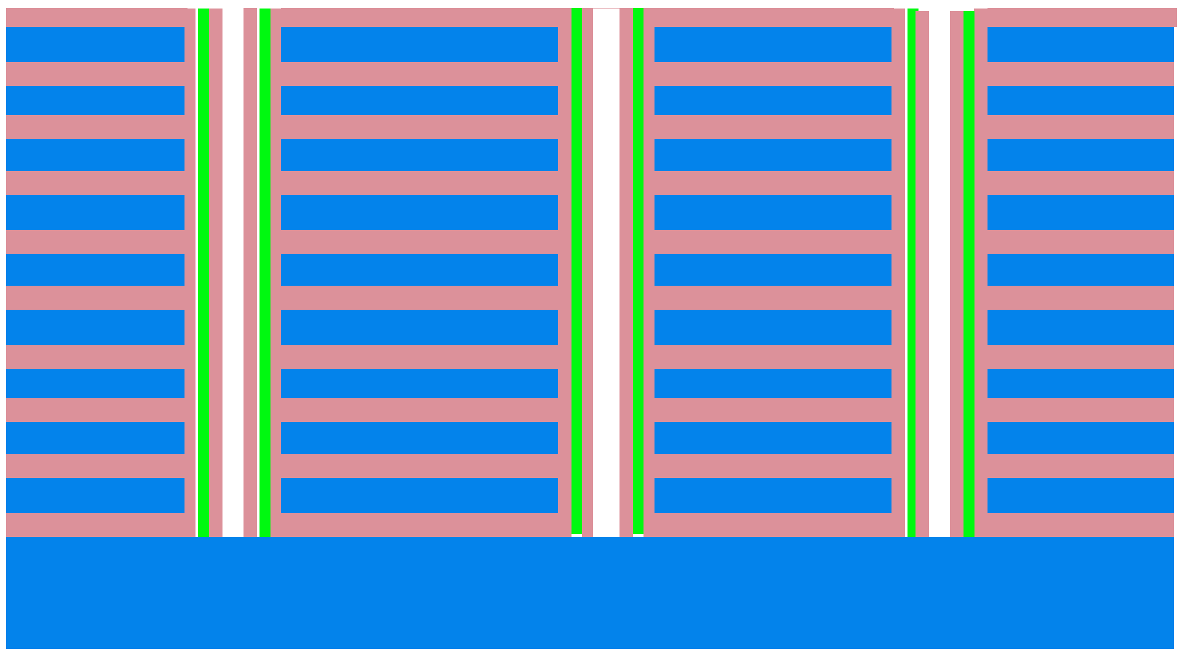
devices shipping in volume



Seagate

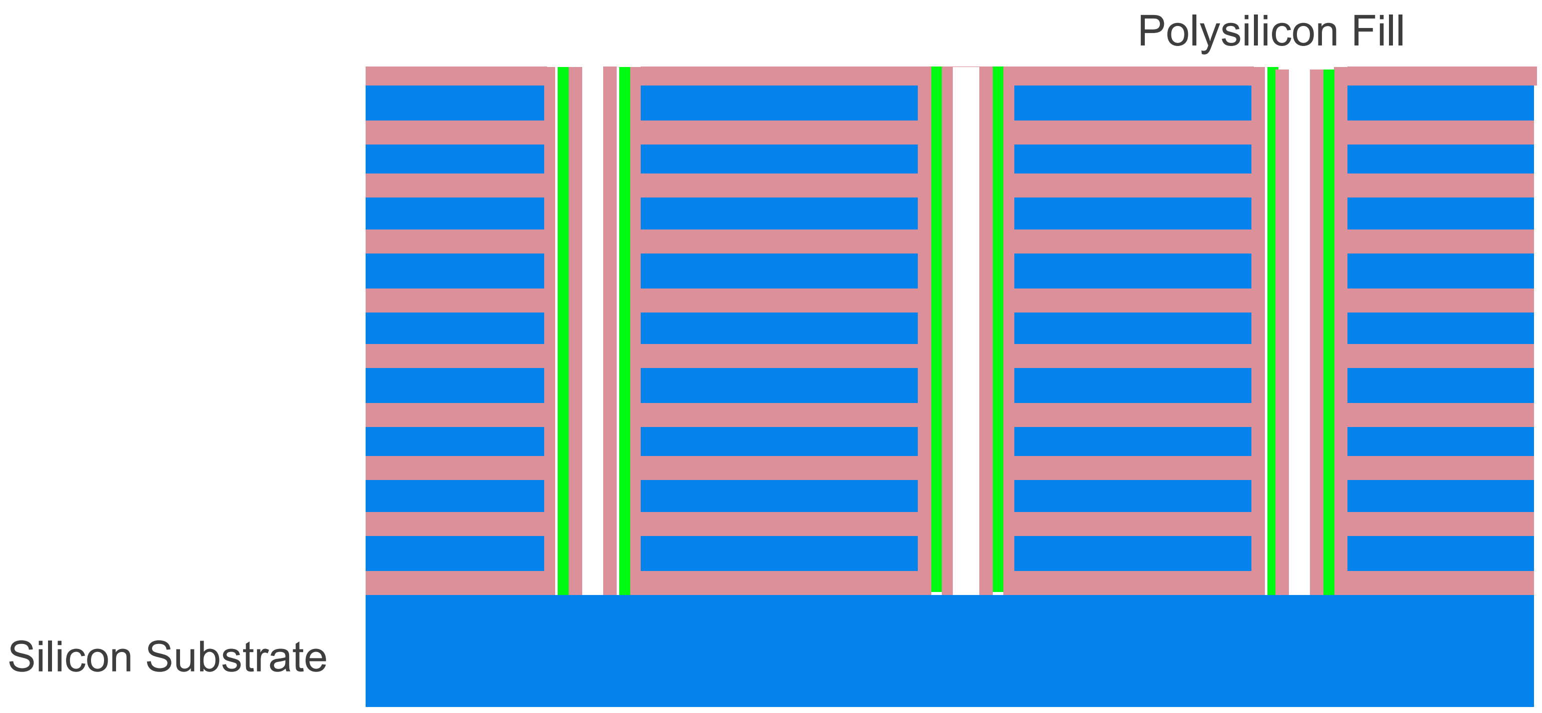
3D NAND Process

devices shipping in volume

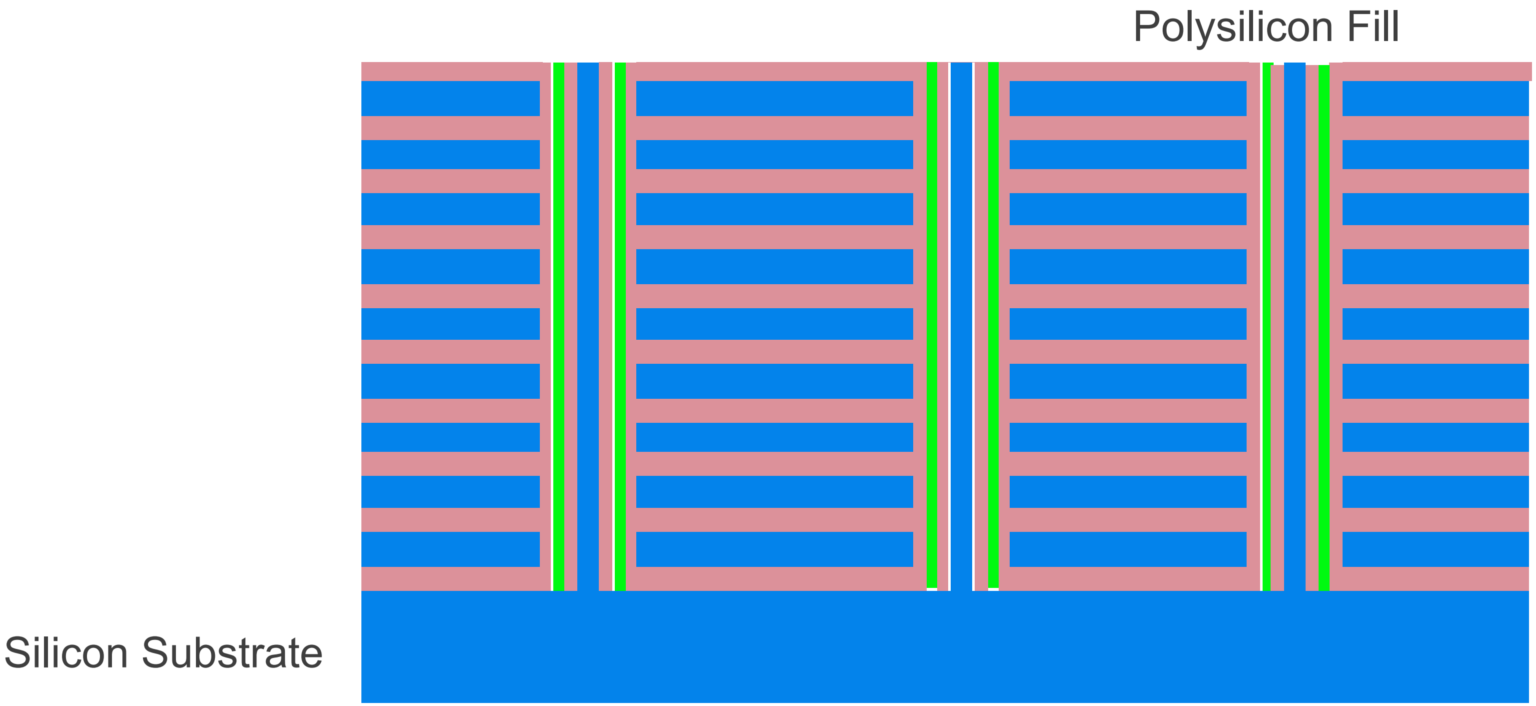


Seagate

3D NAND Process
devices shipping in volume

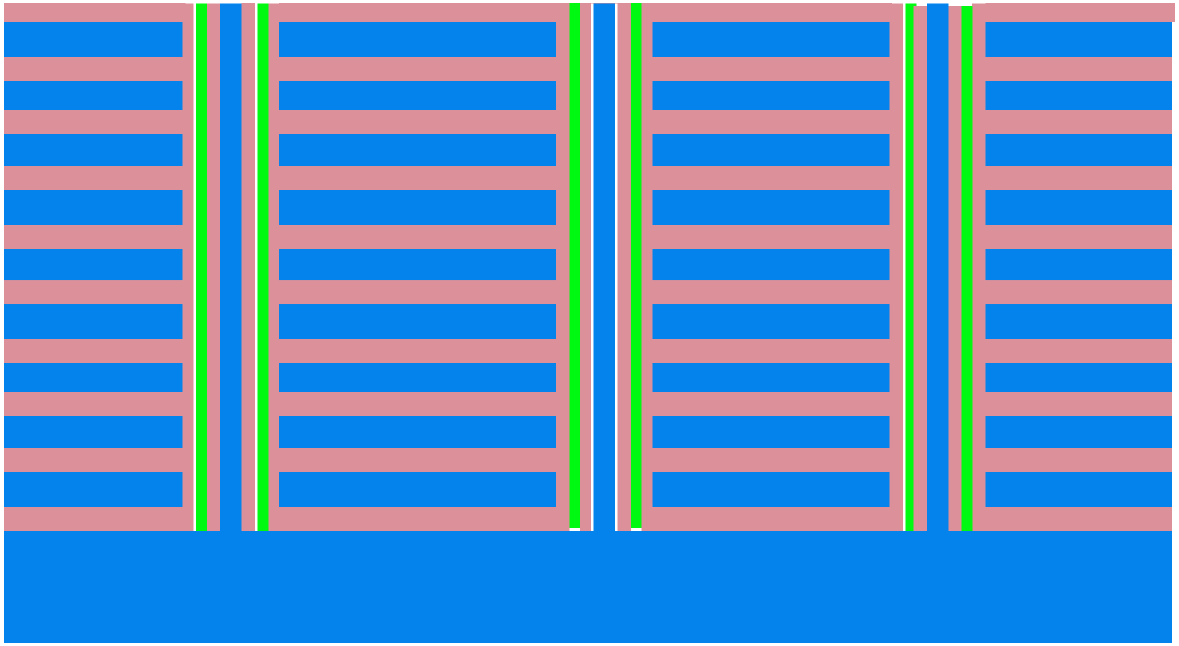


3D NAND Process
devices shipping in volume



3D NAND Process

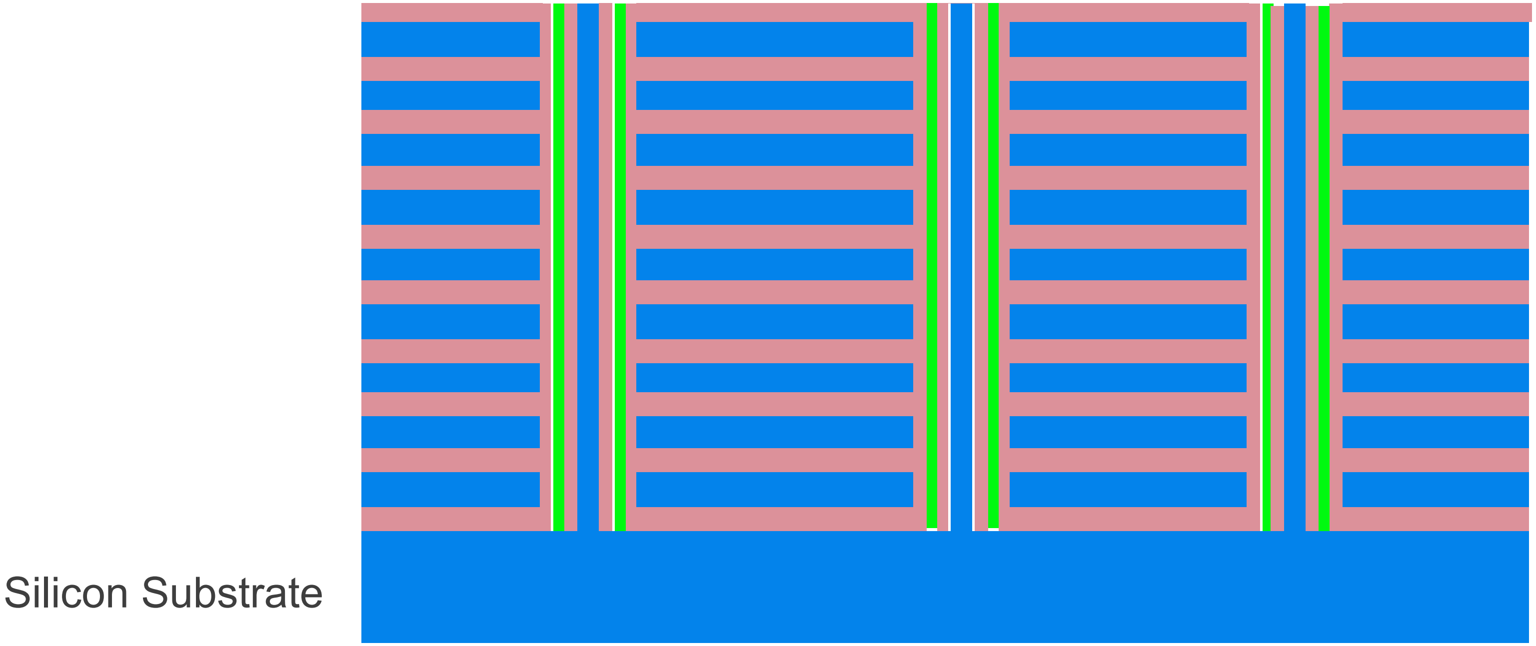
devices shipping in volume



Seagate

3D NAND Process

devices shipping in volume



“The burden will shift from lithography to deposition and etch”
- Ritu Shrivastava, Sandisk

ITRS – Technology Trends

for DRAM and FLASH Memory

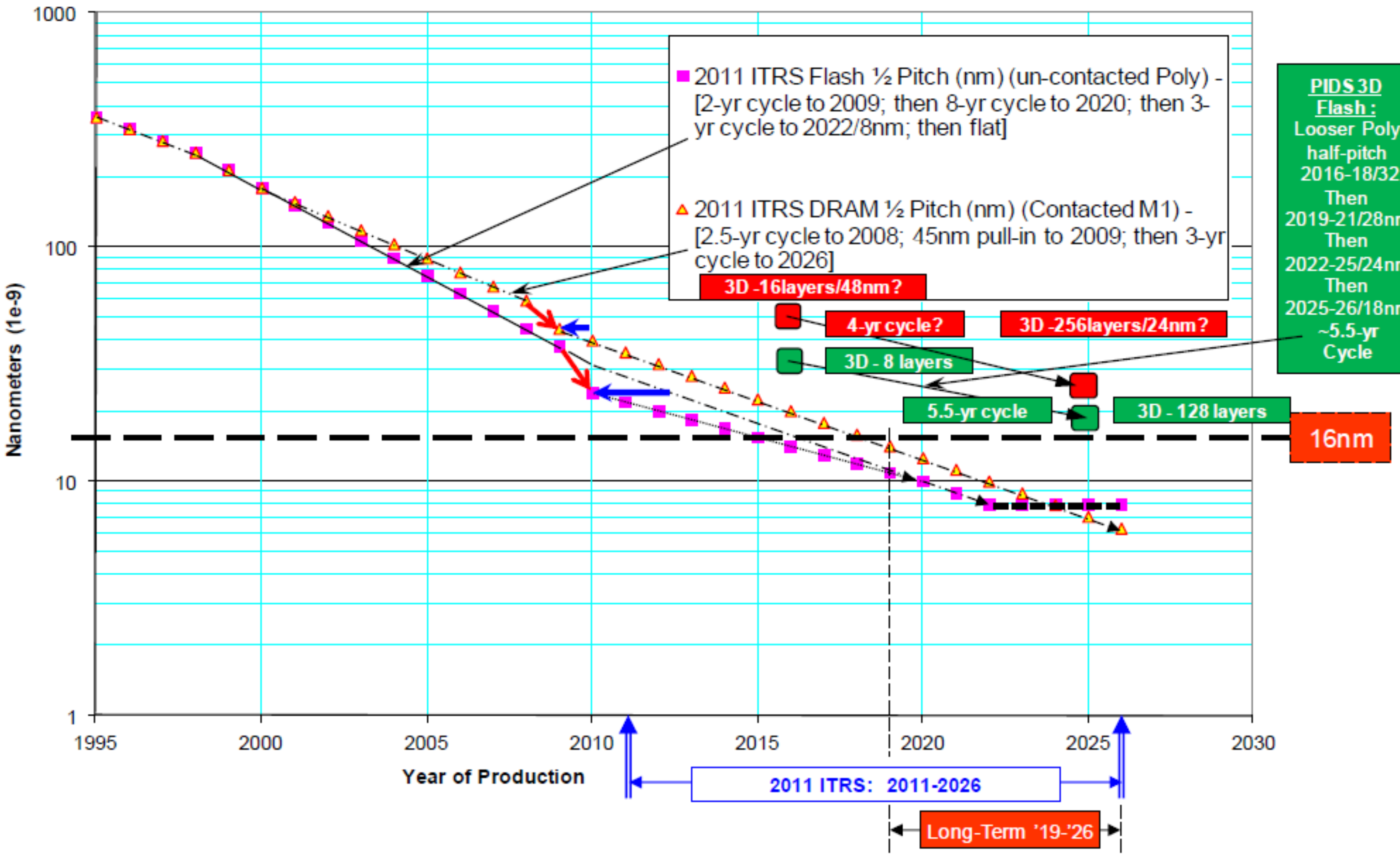


Figure 10 2011 ITRS—DRAM and Flash Memory Half Pitch Trends

Seagate

NAND Flash Scaling - ITRS

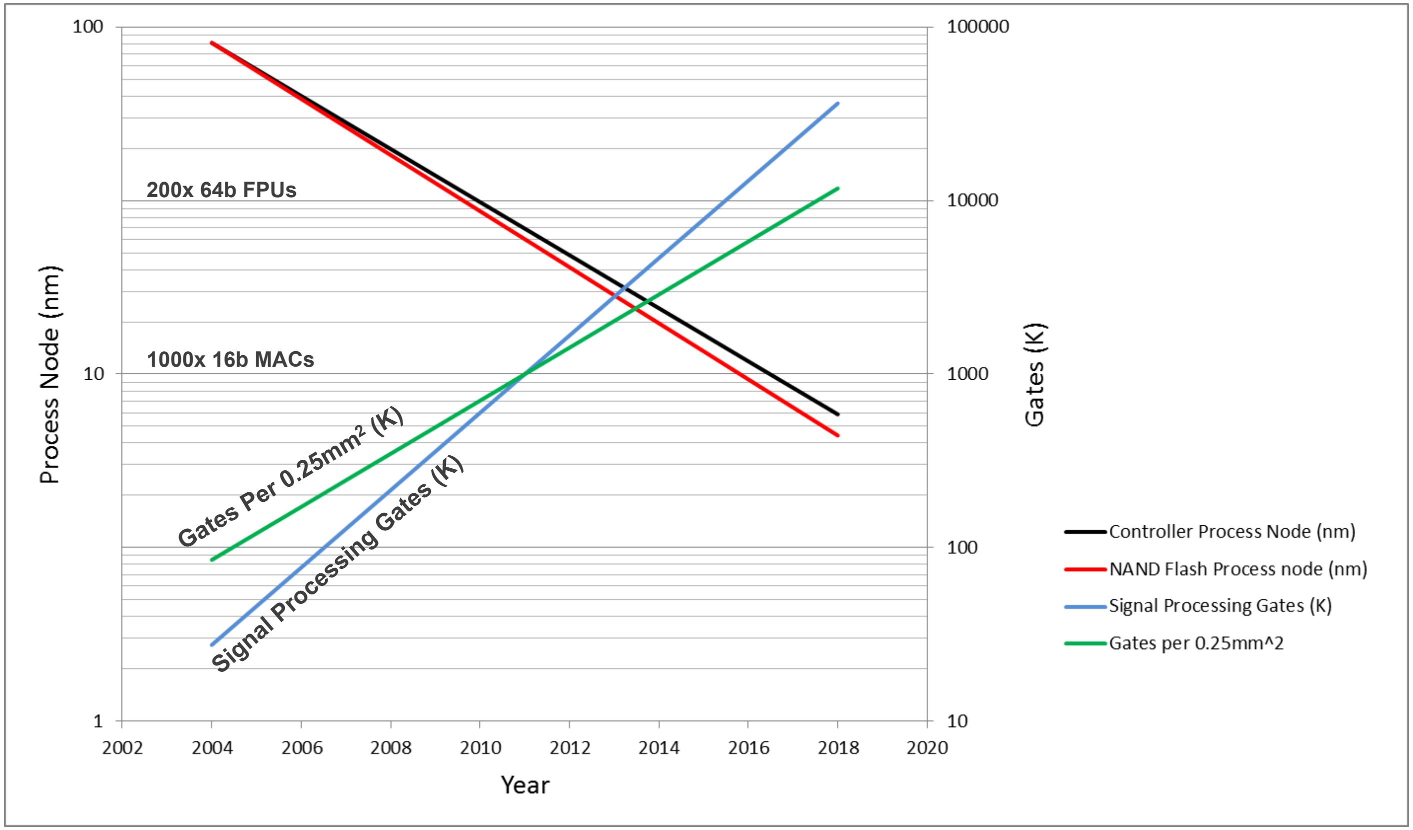
Relax planar scaling, push into 3rd dimension, continue Moore's law

NAND Flash														
Year of Production	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*	*2021*	*2022*	*2023*	*2024*	*2025*
Uncontacted poly 1/2 pitch (nm)	20	18	17	15	14	13	12	11	10	9	8	8	8	8
Number of word lines in one NAND string	64	64	64	64	64	64	64	64	64	64	64	64	64	64
Dominant Cell type	FG	FG	FG/CT	FG/CT	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D	CT-3D
Maximum number of bits per chip (SLC/MLC)					128G / 256G	256G / 512G	256G / 512G	512G / 1T	512G / 1T	512G / 1T	1T / 2T	1T / 2T	1T / 2T	2T / 4T
Minimum array 1/2 pitch - F(nm) [15]					32nm	32nm	32nm	28nm	28nm	28nm	24nm	24nm	24nm	18nm
Number of 3D layers for array at minimum 1/2 array pitch [16]					8	16	32	32	64	64	98	98	98	128

- ITRS Winter Public Conference Dec 2012 Hsinchu, Taiwan

Sustaining NAND density growth

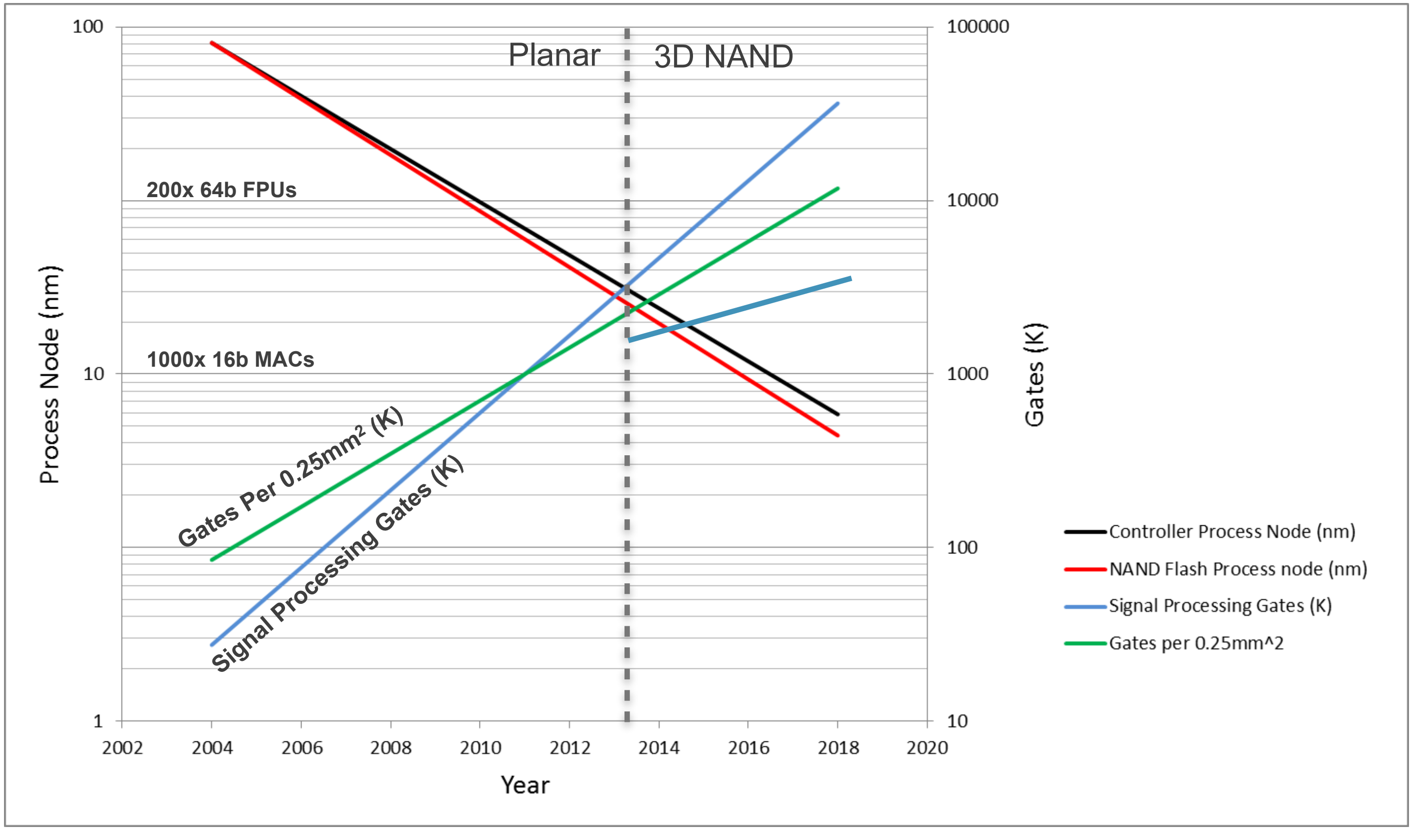
Near Shannon limit, Iterative Low-Density Parity-Check channels >1M gates



Li, Peng, Kevin Gomez, and David J. Lilja. "Exploiting Free Silicon for Energy-Efficient Computing Directly in NAND Flash-based Solid-State Storage Systems." IEEE HPEC 2013

Sustaining NAND density growth

Near Shannon limit, Iterative Low-Density Parity-Check channels >1M gates

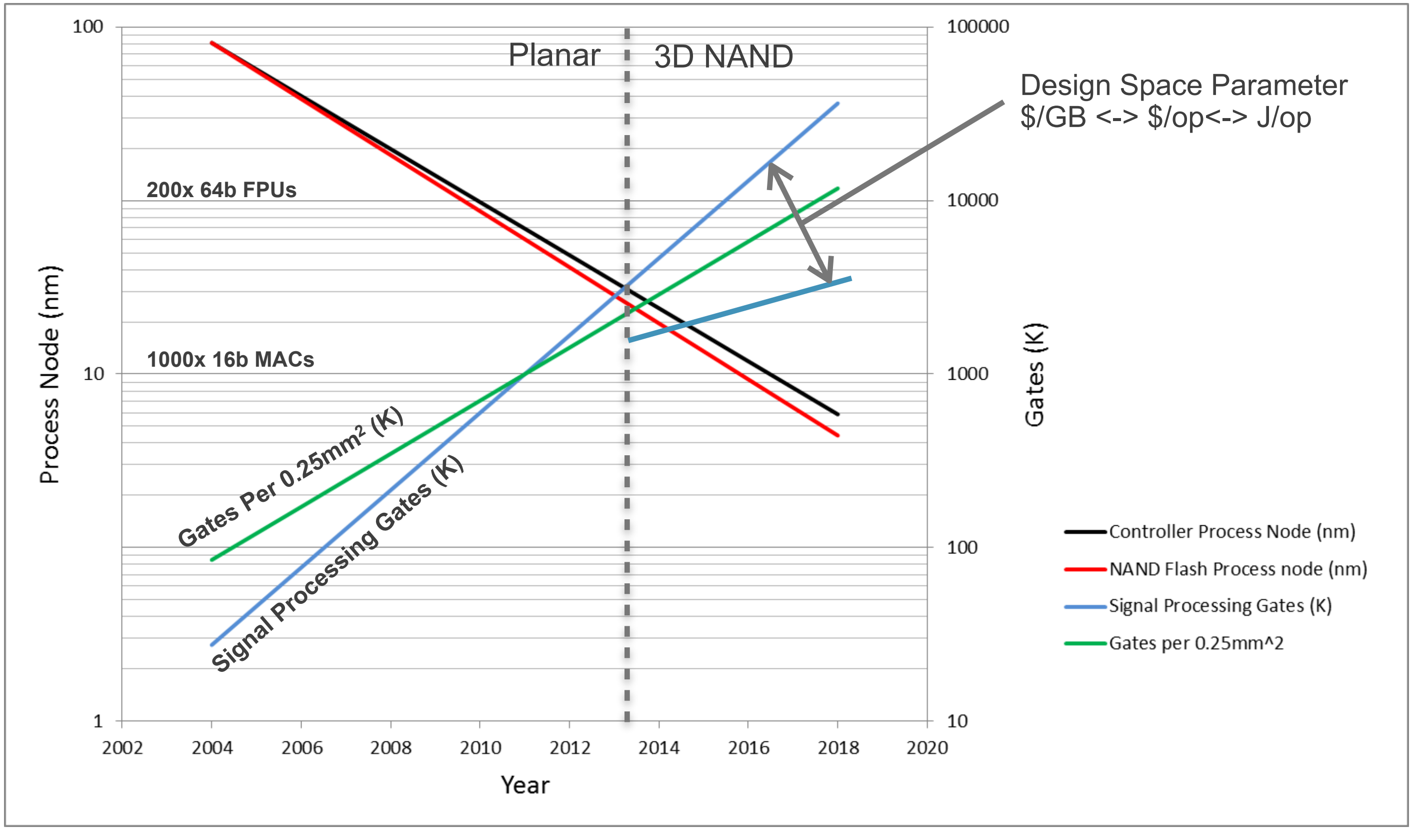


Li, Peng, Kevin Gomez, and David J. Lilja. "Exploiting Free Silicon for Energy-Efficient Computing Directly in NAND Flash-based Solid-State Storage Systems." IEEE HPEC 2013

Seagate

Sustaining NAND density growth

Near Shannon limit, Iterative Low-Density Parity-Check channels >1M gates

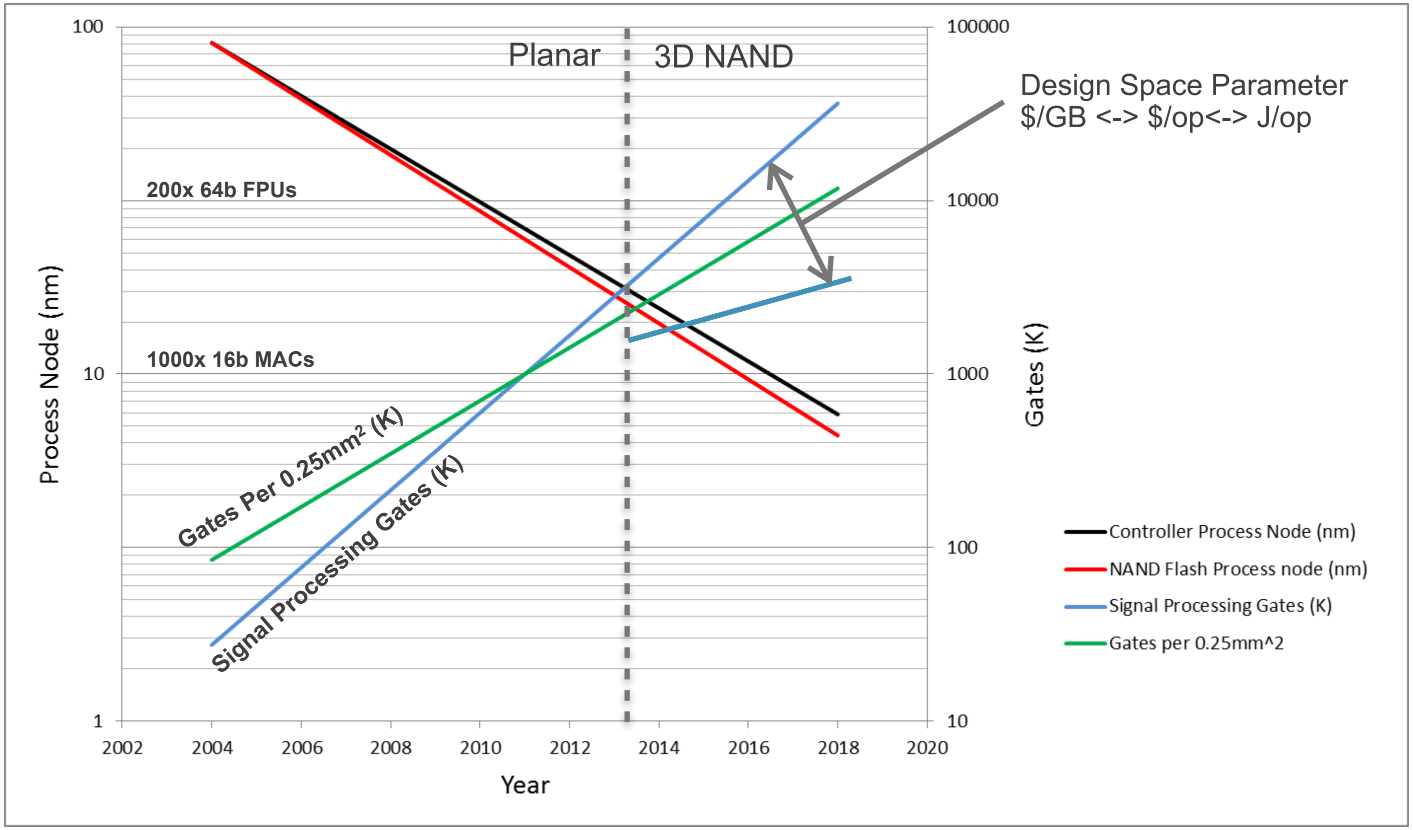


Li, Peng, Kevin Gomez, and David J. Lilja. "Exploiting Free Silicon for Energy-Efficient Computing Directly in NAND Flash-based Solid-State Storage Systems." IEEE HPEC 2013

Seagate

Sustaining NAND density growth

Near Shannon limit, Iterative Low-Density Parity-Check channels >1M gates



Li, Peng, Kevin Gomez, and David J. Lilja. "Exploiting Free Silicon for Energy-Efficient Computing Directly in NAND Flash-based Solid-State Storage Systems." IEEE HPEC 2013

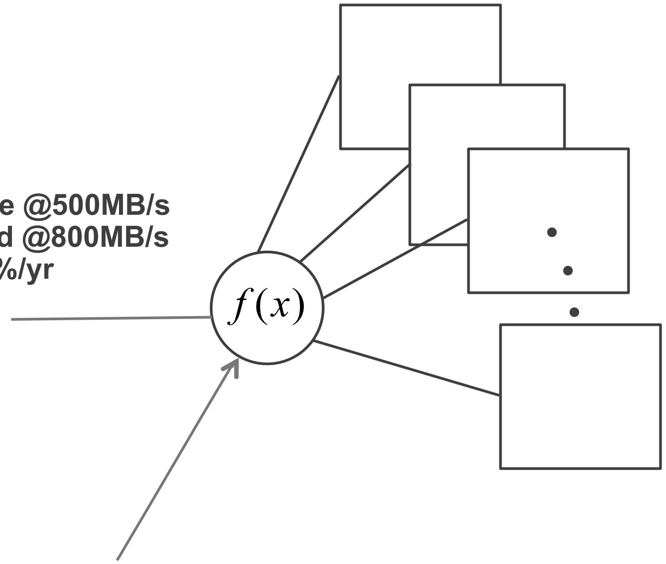
Cost of adding specialized cortical hardware automation is marginal

Seagate

Scaling up a Cortical processor

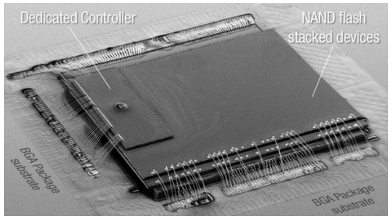
4 to16 Flash Die per Package

Write @500MB/s
Read @800MB/s
+10%/yr

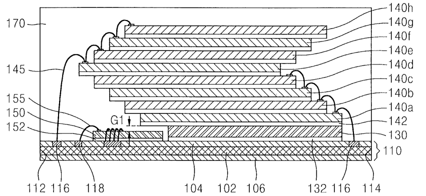


32GB/die +~40%/yr
Write @50MB/s
Read @500MB/s
+20%/yr
Write 20nJ/B
Read 2nJ/B

Currently ~2M gates for LDPC and packet switching
Substrate for specialized or reconfigurable cortical hardware



Example eMMC device (Micron)



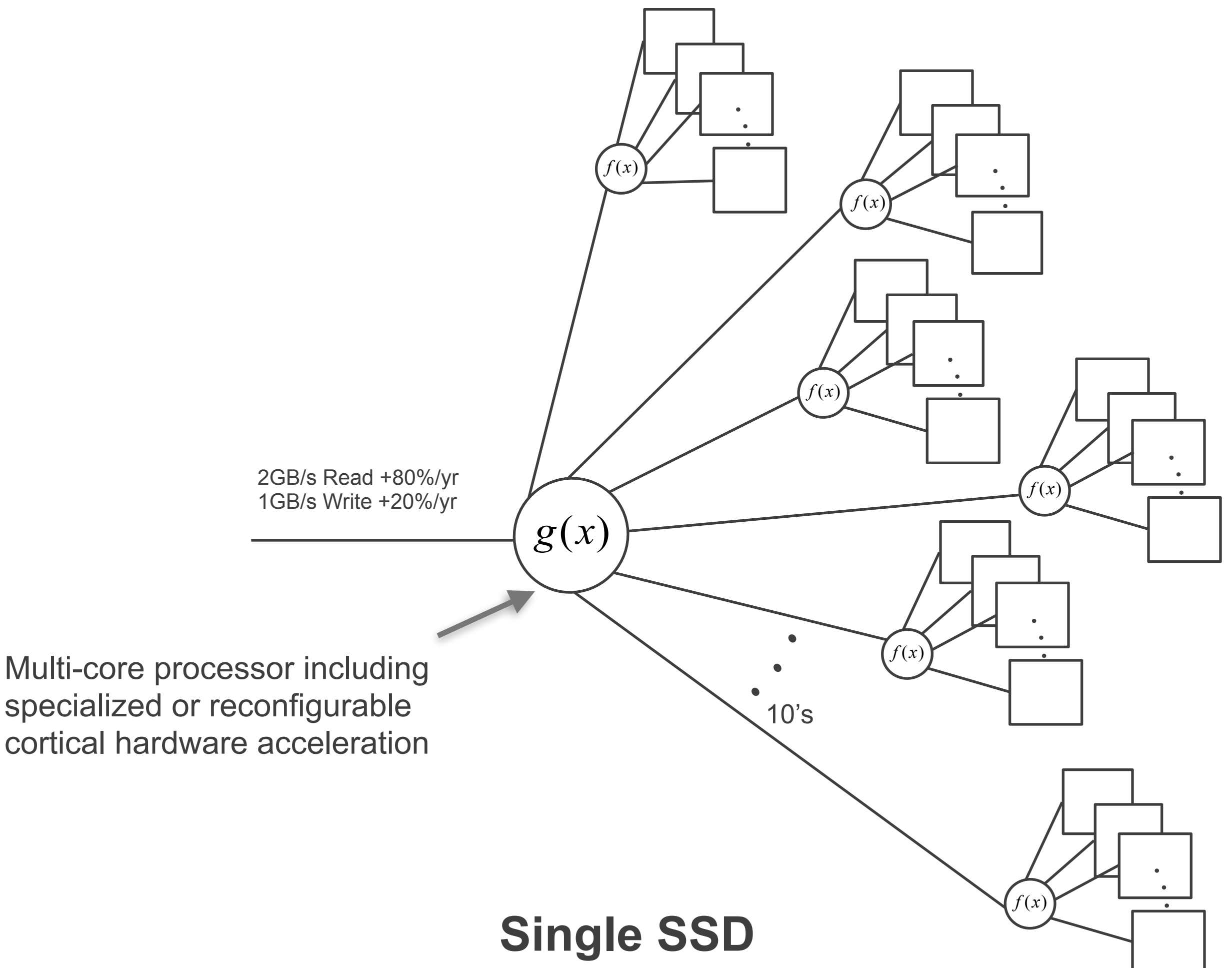
Stacked Die with controller
US 20100314740 A1

Single NAND Flash package ~ 5TB/in^3 +40%/yr

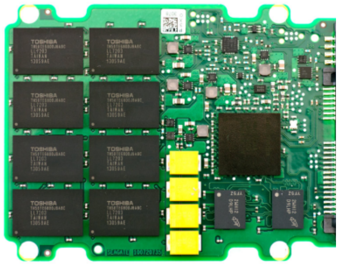
Seagate®

Scaling up a Cortical processor

10’s of Flash Packages in each SSD



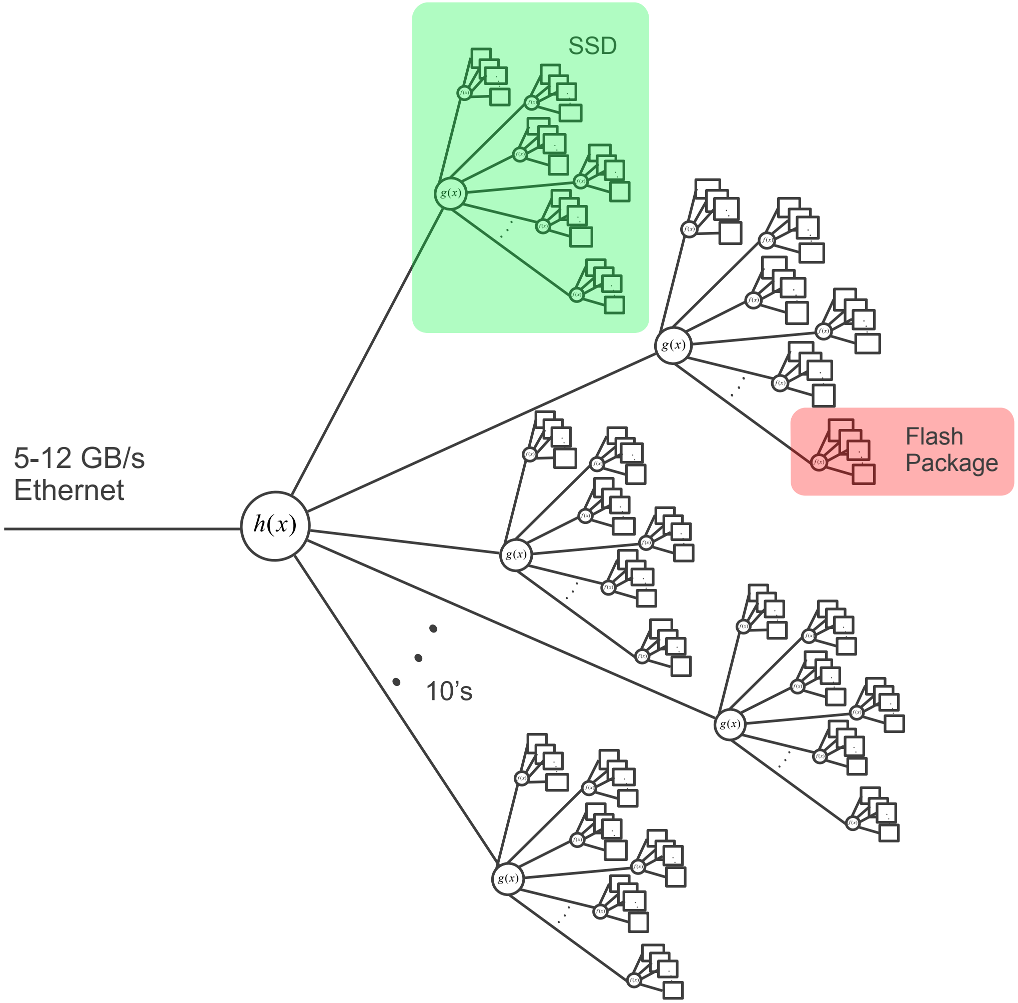
Single SSD
TB’s of NV memory



Seagate 600 Pro SSD

Scaling up a Cortical processor

Array of SSDs in 1U Rack in Cloud Compute Server



Seagate

Convergence of Trends

Why Now

- Flash Memory cost and SNR driven by mobile computing market
  - Increased investment in signal processing silicon at NAND interface – low marginal cost for added compute
- Power Wall – end of Dennard scaling (power $\propto$ 1/L instead of $1/L^3$)
  - Since 2005 has driven multi-core parallelism to maintain compute cost-performance trajectory
  - In turn has forced parallel programming into the mainstream
- Moore's Law post Power Wall continues to provide gates at $1/L^2$ which can not all be switching simultaneously
  - Increased adoption of power islanded heterogeneous architectures operating at device power budget
- Memory Wall – exponentially growing gap between processor and memory performance
  - Continues to drive tighter integration of memory and compute. GPU processing is a temporary reprieve

Heterogeneous Architectures

Lots of efficient H/W automation – powered off most of the time

As Moore continues to increase the number of transistors on silicon at a scale of $1/L^2$ while power is only decreasing as 1/L …

… we can afford to ‘overprovision’ the chip – i.e. use the TDP (total die power budget) using just a subset of the chip’s resources – for example use the entire budget on compute while shutting down global on-chip communication resources.

Enables peak performance (using all available power) on diverse workloads.

This may signal that the right time for Reconfigurable Computing has arrived – specialized hardware acceleration, powered off most of the time.

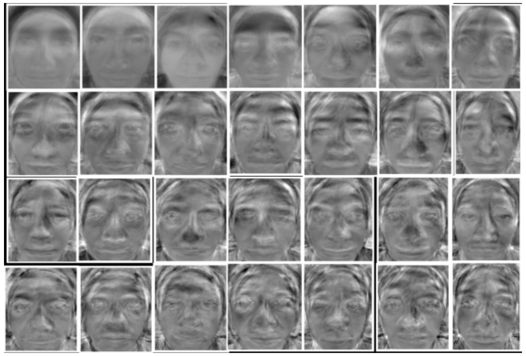
Why NAND Flash and not other NVM technologies

- NAND is a block device and requires a significant and growing investment in signal processing to enable it's continued scaling
- This signal processing overhead is best situated close to NAND to minimize the energy cost of data movement
- NAND has no delusions of being a DRAM replacement like PCM or STTRAM with low-latency and close to byte addressable architectures which will not tolerate any significant signal processing overhead
- **It is not about the technology – it's the economics -** SSDs exist due to the demand for consumer grade NAND devices for the smartphone, tablet, SD Card and USB memory markets.
- Cortical Inspired Compute Elements embedded in SSDs likewise will succeed or fail purely on economics ($/op, J/op) not technology

Seagate®

Architecture Modeling

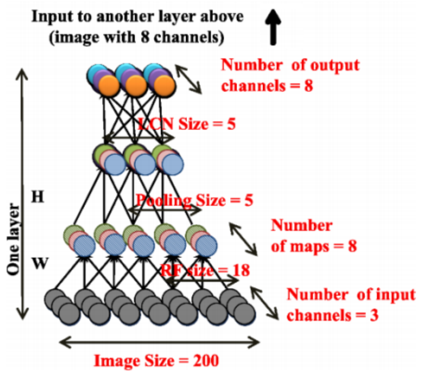
Facial recognition task which is a proxy algorithm for content based image retrieval:
Compute on 16 channel SSD is ~ 0.2mJ/face
150X lower J/face than computing on Host



Brian Harding, Cornell

SCE stores and computes on eigenfaces

Boltzmann machine task- a proxy for many machine learning and data intensive scientific compute algorithms:
Compute on SSD is ~40X lower J/Op compared to Quad-Core host

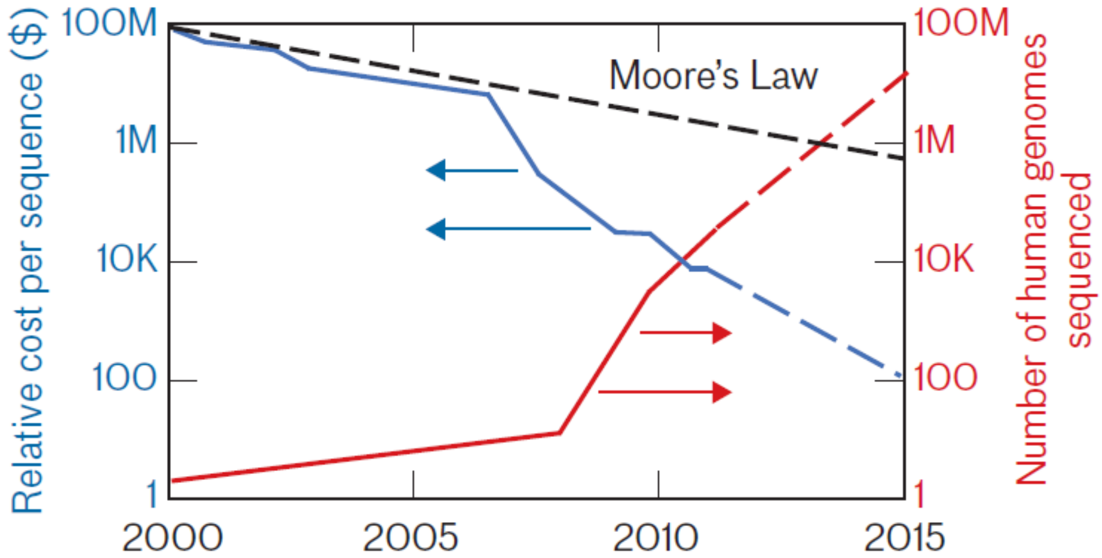


“Building High-level Features Using Large Scale Unsupervised Learning”, Quok Le et al, 2012

Seagate®

The need for Energy Efficiency

Big Data Analytics is no longer a Niche



Advances in DNA sequencing are rapidly decreasing the cost of whole human genome sequencing
As a result, the number of humans being sequenced is increasing significantly
Data needing to be processed is rapidly outpacing computing performance-cost.

Together these drive the need for greater efficiency.

"Taming Biological Data with D4M", Kepner 2013

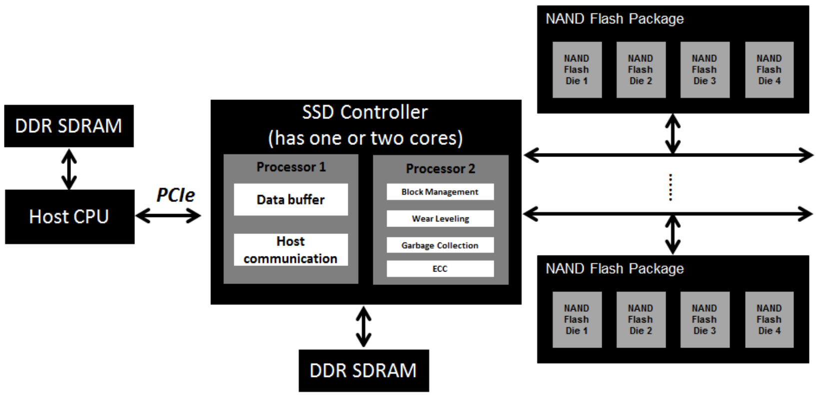
Seagate

Thank You

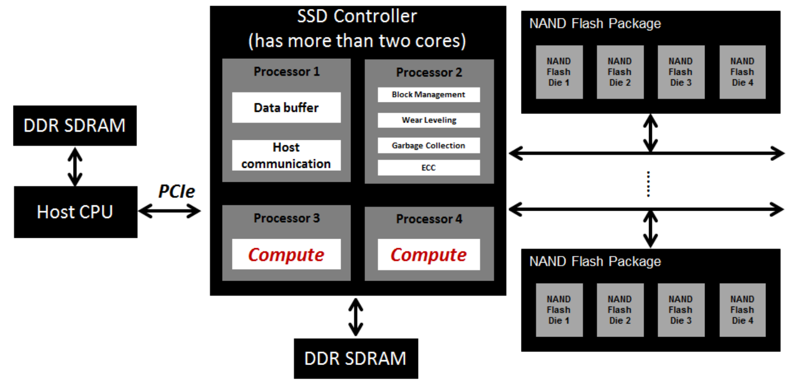
kevin.gomez@seagate.com

Architecture Block Diagrams

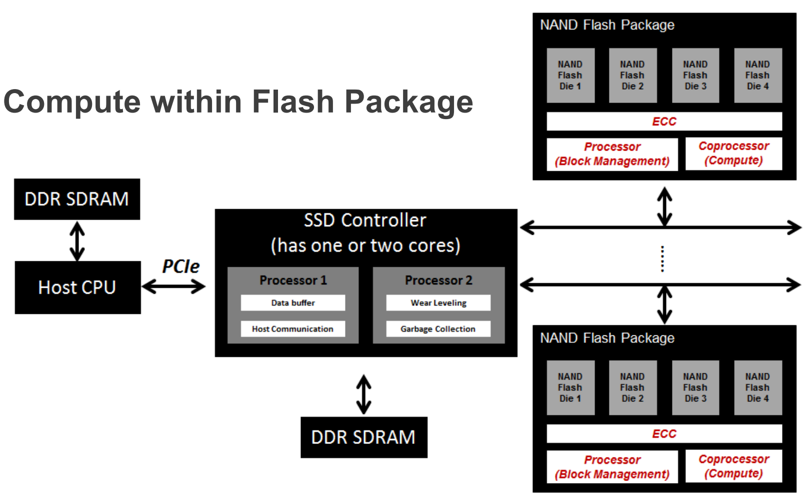
Baseline – SSD for Data, Compute in Host



"Active Flash" – Compute in SSD Controller Processor



Compute within Flash Package



Added Compute functionality may be power islanded and enabled through firmware to make an SCE indistinguishable from a standard SSD - sharing the same production flow and economy of scale

Seagate®

Architecture Simulation Parameters

CPU Power: use ITRS HP technology to evaluate dynamic and leakage power.
Number of Gates: 200M/core
Frequency: 2GHz.
Dynamic Power (per core): 5.04W
Leakage Power (per core): 0.340W

SSD Controller Power: use ITRS LOP technology to evaluate dynamic and leakage power.
Number of Gates: 20 millions per core (Assumption: 10% of the CPU).
Frequency: 1GHz.
Dynamic Power (per core): 0.156W.
Leakage Power (per core): 1.34mW.

Channel Processor Power: use ITRS LOP technology to evaluate dynamic and leakage power.
Number of Gates: 1K, 10K, 100K, 1M.
Frequency: 400MHz.
Dynamic Power (per core): 3.12uW, 31.2uW, 312uW, 3.12mW.
Leakage Power (per core): 67nW, 670nW , 6.7uW, 67uW.

DDR SDRAM: use parameters from MICRON.
Dynamic Power (per 2GB): 438.3mW.
Leakage Power (per 2GB): 88.1mW.

NAND Flash: use parameters from MICRON.
Dynamic Power (per die): 0.04W.
Leakage Power (per die): 0.003W.

Host Interface: PCIe.
Dynamic Power (per GB): 37.5mW.
Leakage Power (per GB): 0.mW

Seagate®

Baseline Face Recognition

	1-Core	2-Core	4-Core	8-Core	16-Core
Average Processing Time of Facial Recognition Algorithm (ms)					
CPI = 100	52.7	26.4	13.3	6.80	3.50
CPI = 10	5.50	2.90	3.40	3.00	2.90
CPI = 1	3.00	2.80	2.80	2.70	2.70
CPI = 0.1	2.70	2.70	2.70	2.70	2.70
Average Power of Facial Recognition Algorithm (W)					
CPI = 100	5.58	10.93	21.5	42.1	81.5
CPI = 10	5.73	10.86	9.97	12.32	15.64
CPI = 1	2.12	2.54	3.27	4.67	7.44
CPI = 0.1	1.39	1.74	2.43	3.81	6.56
Average Energy of Facial Recognition Algorithm (mJ)					
CPI = 100	294	289	286	287	286
CPI = 10	31.5	31.5	33.9	36.9	45.4
CPI = 1	6.36	7.13	9.17	12.6	20.1
CPI = 0.1	3.76	4.71	6.57	10.3	17.7

Core = Host CPU Cores
CPI = clock cycles per instruction of single core in CPU

Active Flash Face Recognition

	1-Core	2-Core	4-Core	8-Core	16-Core
Average Processing Time of Facial Recognition Algorithm (ms)					
CPI = 100	52.6	26.4	13.3	6.70	3.40
CPI = 10	5.40	2.80	1.50	0.800	0.500
CPI = 1	0.700	0.400	0.300	0.300	0.300
Average Power of Facial Recognition Algorithm (W)					
CPI = 100	0.699	0.858	1.17	1.79	2.98
CPI = 10	0.716	0.881	1.18	1.70	2.48
CPI = 1	0.839	1.02	1.15	1.16	1.17
Average Energy of Facial Recognition Algorithm (mJ)					
CPI = 100	36.8	22.6	15.6	12.0	10.1
CPI = 10	3.86	2.47	1.78	1.36	1.24
CPI = 1	0.587	0.410	0.345	0.347	0.351

Core = SSD Controller Cores
CPI = clock cycles per instruction of single core in SSD controller

Exploiting Free Silicon for Energy-Efficient Computing Directly in NAND Flash-based Solid-State Storage Systems, High Performance Extreme Computing 2013, Li et al

In-Flash-Package Face Recognition

Channels	4	8	16	32
Average Processing Time (ms)				
Time	0.300	0.200	0.100	0.0500
Average Power of Facial Recognition Algorithm (W)				
Gates = 1K	0.887	1.23	1.87	2.98
Gates = 10K	0.887	1.23	1.87	2.98
Gates = 100K	0.888	1.23	1.88	2.99
Gates = 1M	0.899	1.26	1.92	3.06
Average Energy of Facial Recognition Algorithm (mJ)				
Gates = 1K	0.266	0.246	0.187	0.149
Gates = 10K	0.266	0.246	0.187	0.149
Gates = 100K	0.266	0.247	0.188	0.149
Gates = 1M	0.270	0.251	0.192	0.153

References

[1] “Hitting the memory wall: implications of the obvious”, WA Wulf, SA McKee - ACM SIGARCH computer architecture news, 1995

[2] “Reflections on the memory wall”, SA McKee - Proceedings of the 1st conference on Computing, 2004

[3] “Missing the memory wall: The case for processor/memory integration”, A Nowatzyk, F Pong, A Saulsbury - Architecture, 1996

[4] “Computing performance: Game over or next level?”, SH Fuller, LI Millett - Computer, 2011

[5] “Platform 2015: Intel processor and platform evolution for the next decade”, S Borkar, P Dubey, K Kahn, D Kuck, H Mulder

[6] “Dark silicon and the end of multicore scaling”, H Esmaeilzadeh, E Blem, RS Amant… (ISCA), 2011

[7] “GPUs and the future of parallel computing”, SW Keckler, WJ Dally, B Khailany, M Garland - Micro, 2011

[8] “The GPU computing era”, J Nickolls, WJ Dally - Micro, IEEE, 2010

[9] “Architecture at the End of Moore”, S Kaxiras - 2013 – Springer

[10] “The Shift to Cloud Computing: Forget the Technology, It’s About Economics”, Jim Cooke 2010

[11] “An Energy-Efficient Processor Architecture for Embedded Systems”, Balfour et. al. 2008

[12] “Trends in Computation, Communication and Storage and the Consequences for Data-intensive Science”, Oliveira, S.F., 2012

Seagate